Estimação em Estratos Sub-representados no Contexto das Sondagens Eleitorais - Uma Comparação de Métodos

Por

João Filipe Abelha Figueiredo

Tese de Mestrado em

Análise de Dados e Sistemas de Apoio à Decisão

Orientada por:

Professor Doutor Pedro Campos

2010

Nota Biográfica do Candidato

João Filipe Abelha Figueiredo nasceu em Clermont Ferrand, França em 1982, vivendo contudo em Portugal desde o ano de 1983. Licenciou-se em Matemática, ramo Científico de Matemática Aplicada em 2006 pela Faculdade de Ciências da Universidade do Porto. Em 2006 inscreveu-se na Especialização em Estatística Aplicada e Modelação, na Faculdade de Engenharia da Universidade do Porto, não a tendo contudo concluído. Começou a trabalhar em 2007 na área da análise e tratamento de dados e como formador de estatística e matemática, funções que desempenha até à data, cumulativamente com a função de editor de conteúdos multimédia para o ensino básico. Em 2007 inscreveu-se no Mestrado em Análise de Dados e Sistemas de Apoio á Decisão, na Faculdade de Economia da Universidade do Porto, tendo terminado a parte curricular com média de 14 valores. Desde 2010 frequenta a Licenciatura em Engenharia Mecânica, no Instituto Superior de Engenharia do Porto.

Agradecimentos

Esta página é para todos aqueles que tornaram possível este trabalho. Ao Professor Doutor Pedro Campos pela confiança depositada, pelo apoio e disponibilidade demonstrada e pela orientação neste trabalho. Aos meus pais, aos meus avós e a toda a minha família, todos sempre presentes mesmo que alguns distantes, pelo incentivo e confiança permanentemente demonstrados. À Nathalie por nunca ter duvidado. À Leonor e a todas as crianças pelos sorrisos e pela alegria. Ao Professor Doutor Aguiar Falcão pela disponibilização dos dados e pela cooperação. A todos os amigos pela amizade e confiança.

À Ana pelo Amor, pelo apoio, pela partilha, por ser quem é, por me fazer ser melhor, por tudo…

E a todos,

O meu obrigado.

Este trabalho é vosso.

Resumo

As sondagens são fundamentais como instrumento de recolha de informação. Actualmente, e cada vez mais, a existência de informação é uma vantagem competitiva que pode fazer a diferença. No caso das eleições legislativas em Portugal a eleição dos deputados é feita a nível de distrito. Tome-se uma sondagem elaborada com o objectivo de averiguar a intenção de voto a nível de Portugal continental, com uma amostra obtida pelo método de amostragem aleatório estratificado. Como uma amostra representativa da opinião a nível de Portugal continental (nível macro) não o é necessariamente a nível de distrito (nível micro), neste trabalho procura-se optimizar a utilidade de uma sondagem eleitoral, procurando melhorar as estimativas num dado partido, quer a nível macro, quer a nível micro, através da aplicação de três metodologias distintas: (i) agregação de sondagens, fazendo uso de eventuais sondagens auxiliares existentes, (ii) aplicação de métodos de regressão multinível, fazendo uso da estrutura multinível dos dados em estudo, (iii) métodos de estimação em pequenos domínios, fazendo uso de eventuais informações secundárias, através da aplicação do Empirical Best Linear Unbiased Prediction (EBLUP). Estes métodos foram aplicados a dados reais, provenientes de uma sondagem realizada para estimar a intenção de voto nas eleições legislativas de 2005. O objectivo definido consistiu na melhoria das previsões de voto no PS. Dos modelos estudados, aquele que evidenciou melhores resultados foi o EBLUP.

Abstract

The polls are crucial as a tool for gathering information. Currently, and increasingly, the availability of information is a competitive advantage that can make a difference. In the case of the legislative elections in Portugal the election of deputies is made at the district level. Let's consider a survey conducted with the aim of ascertaining the intention of voting at the level of Portugal, with a sample obtained from a stratified random sampling. As a representative sample of opinion in terms of mainland Portugal (macro level) is not necessarily representative at the district level (micro level), this work seeks to optimize the utility of an election poll, trying to improve estimates in a given party, both at macro and micro level, by applying three different methodologies: (i) aggregation of surveys, using any existing auxiliary polling, (ii) application of multilevel regression methods, using the multilevel structure of the data in study, (iii) estimation methods in small areas, making use of any secondary information, through the application of Best Linear Unbiased Prediction Empirical (EBLUP). These methods were applied to real data, taken from a survey conducted to estimate the intention to vote in 2005 legislative elections. The target set was the improving of the predictions of voting in the PS. Of the models studied, the one that showed better results was EBLUP.

Índice

Índice de Gráficos

Índice de Tabelas

Índice de Figuras

Capítulo 1 - Introdução

Neste capítulo faz-se a apresentação do problema em estudo e uma pequena contextualização dos métodos aplicados e estudados ao longo deste trabalho, bem como das principais conclusões obtidas.

As sondagens são uma ferramenta útil e preciosa para a obtenção de informação sobre os mais diversos temas e áreas, seja nas áreas da sociologia, marketing, económicas ou políticas. Numa sociedade, do conhecimento, informação é sinónimo de vantagem competitiva para quem a detém. Segundo Reis et al (2001), a sondagem é hoje, em vários domínios, a resposta ao conhecimento de uma população tomando por base uma sua fracção – a amostra. E, para Cochran (1977), os nossos conhecimentos, atitudes e acções são baseadas, em grande medida, em amostras. Sendo que isto é verdade quer nos assuntos do dia-a-dia, quer na investigação científica.

No caso das sondagens eleitorais, a informação é determinante para os cidadãos e para os políticos, uma vez que lida com o futuro e com o bem comum. A informação proveniente das sondagens eleitorais é importante quer para os candidatos (no sentido de uma boa preparação das campanhas, temas a abordar, averiguação de verdadeiras hipóteses de eleição), quer para os eleitores (para uma escolha acertada dos candidatos).

Se informação é sinónimo de vantagem, uma sondagem implica custos. Daí a necessidade de optimizar a relação informação/custo ser hoje uma questão tão importante quanto a própria obtenção de informação. Para a diminuição do custo procura-se diminuir o tamanho das amostras, o que acarreta problemas ao nível da qualidade da informação. Em alternativa pode-se recolher informação diferenciada na mesma sondagem, por exemplo através do aumento do número de questões no mesmo questionário. Contudo, esse procedimento torna o processo de recolha mais extenso e com fraca aderência à participação dos respondentes.

Um dos dilemas actuais das sondagens, em particular das sondagens eleitorais reside na dificuldade de estimação da proporção de votantes num determinado partido em situações em que a amostra é de dimensões reduzidas. No caso de uma sondagem eleitoral estamos interessados em estimar a proporção de eleitores, p, que, a nível nacional, votará num determinado partido, assumindo que a população em estudo se encontra dividida em L estratos, por exemplo distritos. Tome-se um partido A. Neste trabalho procura-se melhorar as estimativas da proporção de votos nesse partido obtida através de uma sondagem eleitoral comparando três abordagens diferentes:

(i) Agregação de sondagens, através da aplicação de uma variação do trabalho de Erikson et al (1993) que, de modo a estimar a opinião pública a nível nacional nos EUA, procedeu à agregação de 122 sondagens da CBS/NYT (Columbia Broadcasting System News/New York Times), todas realizadas com o mesmo alcance geográfico, segundo a mesma base e mantendo as mesmas questões.

(ii) Aplicação de métodos de regressão multinível, seguindo o trabalho de Gelman et al (1997) e (2007) e Park et al (2004) que procederam à aplicação de modelos de regressão multinível em dados provenientes de sondagens eleitorais realizadas pela CBS/NYT para a obtenção de melhores estimativas da proporção de votos nas eleições Presidenciais nos EUA de 1988 a nível de estado. Estas estimativas foram depois pós-estratificadas de modo a corrigir as diferenças existentes entre a distribuição da amostra e a distribuição da população.

(iii) Utilização de métodos de estimação em pequenos domínios, aplicando o estimador EBLUP (Empirical Best Linear Unbiased Prediction) a nível área (Rao (2003) e fazendo uso de informação secundária.

O objectivo deste trabalho consiste na comparação dos resultados provenientes da aplicação destas técnicas, o que constitui algo original, pois estas técnicas nunca foram comparadas e aplicadas em simultâneo no contexto da estimação através de sondagens eleitorais. Através da aplicação destas técnicas procura-se, não só melhorar as estimativas a nível global, possibilitando a utilização de amostras mais pequenas, mas também melhorar as estimativas a outros níveis de desagregação dos dados,

nomeadamente a nível de distrito, possibilitando, no limite, a dispensa da elaboração de uma nova sondagem.

Para a comparação destas técnicas, fixou-se o problema na estimação da proporção de votos num dado partido (por exemplo, PS) e procedeu-se à aplicação das metodologias numa sondagem, realizada pelo **IPOM** – Instituto de Pesquisa de Opinião e Mercado, no ano de 2005 para averiguar a intenção de voto nas eleições legislativas desse ano e com um alcance regional a nível de Portugal Continental. A obtenção de melhores estimativas a nível de distrito é importante no caso das eleições para a Assembleia da República uma vez que a eleição dos deputados é feita a nível de distrito. Uma sondagem a nível nacional poderá ter um tamanho de amostra suficiente para dar uma boa estimativa da proporção global de votos num dado partido e, contudo, não possuir tamanhos de amostra suficientes a nível dos distritos para fornecer boas estimativas das proporções de votos a esse nível e, consequentemente, do número de deputados a eleger por distrito.

Dado que através do método de agregação de sondagens se procede, como o próprio nome indica, ao agrupamento de sondagens, a inexistência de uma maior quantidade de sondagens auxiliares impossibilitou um maior aprofundamento deste método. Este é um método que deve ser aplicado com precaução, pois, uma vez que a intenção de voto varia com o tempo, ao fazer o agrupamento de sondagens devemos garantir que estas foram elaboradas no mesmo espaço temporal e que nenhum acontecimento decisivo na definição da intenção de voto (por exemplo um debate entre os candidatos) tenha sucedido entre a elaboração das sondagens. Com esta metodologia pretende-se melhorar as estimativas, quer a nível macro (continente), quer a nível micro (distrito), fazendo recurso do aumento do tamanho de amostra proveniente do agrupamento das sondagens. Uma vez que não se dispunham de sondagens auxiliares realizadas com o mesmo alcance geográfico, aplicou-se uma variação do método utilizado por Erikson et al (1993), na medida em que se recorreu a duas sondagens com um alcance geográfico a nível micro, nomeadamente a nível dos distritos de Aveiro e do Porto, para agrupar com a sondagem a nível continental. O aumento do tamanho de amostra nos distritos afectados pelo agrupamento permitiu uma melhoria considerável nas respectivas

previsões (e, consequentemente, uma ligeira diminuição no erro absoluto médio). As previsões a nível de distrito foram pós-estratificadas, de modo a estarem de acordo com a distribuição populacional, e obteve-se desta forma a previsão a nível macro, que por sua vez é pior do que a previsão fornecida exclusivamente pela sondagem tomada isoladamente. O excelente resultado obtido a nível do distrito do Porto dá a entender que este método poderá ser encarado como uma alternativa válida a considerar, quando existirem condições logísticas que o permitam, para a melhoria das estimativas fornecidas por uma sondagem.

Fazendo uso das variáveis demográficas existentes nos dados obtidos através da sondagem foram definidos quatro modelos multinível, cujos parâmetros foram simulados fazendo recurso do WinBUGS (Bayesian inference Using Gibbs Sampling) (Lunn (2000)). As simulações obtidas pelo WinBUGS foram de seguida pós-estratificadas (de modo a corrigir eventuais divergências entre a distribuição da amostra e a distribuição populacional no que se refere às variáveis demográficas) e sumariadas num ponto preditor (a média) obtendo desta forma a previsão final da proporção de votos num dado partido. Através da definição de diferentes modelos procurou-se encontrar um modelo que tivesse um bom ajuste ao conjunto de dados a modelar e que fornecesse boas estimativas das proporções de votos no partido escolhido, a nível macro e a nível micro. O modelo de regressão multinível que obteve os melhores resultados foi definido utilizado parâmetros redundantes (de modo a aumentar a velocidade de convergência Gelman et al (2007)) e faz recurso de todas as variáveis demográficas, bem como de algumas interacções entre as variáveis (considerando como declives as variáveis sexo, profissão e a interacção entre elas e como coordenadas na origem as variáveis idade, escolaridade, interacção entre as variáveis idade e escolaridade e distrito). O modelo definido nestes termos apresenta melhorias substanciais nas previsões dos distritos com menor tamanho de amostra, face às previsões consideradas tomando isoladamente a sondagem. Este facto permitiu uma redução notória do erro absoluto médio das previsões. Foi também possível verificar que, apesar de apresentarem resultados relativamente semelhantes a nível continental, a previsão fornecida pelo modelo multinível está mais próxima dos resultados reais verificados.

A ideia original para a aplicação do estimador EBLUP era a utilização de diferentes sondagens como informação auxiliar, tal não foi no entanto possível devido à não acessibilidade/disponibilidade dessa informação. Como tal, procedeu-se à definição de dois modelos EBLUP a nível de área, utilizando como informação auxiliar os resultados eleitorais das eleições legislativas de 2002 e de 1999. Os resultados obtidos permitem concluir que a escolha da informação auxiliar, aquando da aplicação deste método, é de extrema importância e possui um impacto fundamental nos resultados fornecidos pelo método. Dos dois métodos definidos, o que obtém melhores resultados é aquele que faz recurso da informação auxiliar relativa ao ano de 1999, tal sucede devido ao facto da intenção de voto em 2005 ser mais próxima dos resultados eleitorais de 1999 do que dos de 2002. Devido à existência desta proximidade, no que se refere à tendência de voto entre 2005 e 1999, este modelo é, de todos, o que apresenta melhores resultados no que se refere às previsões a nível de distrito, especialmente nos distritos com menor peso amostral. A previsão a nível macro obtida por este método apresenta uma ligeira melhoria relativamente à previsão fornecida pela sondagem.

Esta tese encontra-se estruturada da seguinte forma: no Capítulo 2 faz-se a definição do problema e a apresentação e formalização das metodologias aplicadas e estudadas; no Capítulo 3 apresentam-se os resultados obtidos; e no Capítulo 4 apresentam-se as conclusões obtidas neste trabalho.

Capítulo 2 - Alguns Métodos para Estimação em Estratos Sub-representados em Sondagens Eleitorais

Nesta secção faz-se a contextualização do problema em estudo: o da estimação em sondagens eleitorais com situações de sub-representação de estratos, ou seja em situações em que as amostras de alguns estratos apresentam dimensões reduzidas. Para além disso, apresentam-se três tipos de métodos até agora encarados como alternativos para a estimação da proporção de votantes nessas situações: Estimação em Pequenos Domínios, Regressão Multinível e Agregação de Sondagens.

2.1 – A Estimação da Proporção de Votantes num Determinado Partido

Um dos dilemas actuais das sondagens, em particular das sondagens eleitorais reside na dificuldade de estimação da proporção de votantes num determinado partido em situações em que a amostra tem dimensões reduzidas. Tome-se o exemplo de pequenos distritos, ou municípios, com representação reduzida no espaço eleitoral nacional. Nessas situações, existem problemas de estimação relacionados com elevada variabilidade das estimativas produzidas que podem conduzir a grandes diferenças entre os resultados da sondagem e os resultados observados na população.

Em geral, as sondagens são capazes de fornecer uma boa estimativa para o parâmetro em estudo na população, apesar de estarem associadas a certos tipos de erros: (i) erros amostrais provenientes do facto de se estar a inquirir apenas parte da população e não toda a população, que serão importantes no decorrer deste estudo e (ii) erros não amostrais, relacionados com eventuais erros que possam suceder no processo de recolha e tratamento da informação. [1]

[1] Uma quantificação melhor dos tipos erros não amostrais de erros pode encontrar-se em Cochran (1977), que subdivide os erros do tipo (ii) em: (a) erros oriundos da impossibilidade de inquirir algumas das unidades escolhidas para a amostra, (b) erros na medida da opinião, e (c) erros na introdução dos dados.

A sondagem pode ser vista como um processo geral da consulta da população e a amostragem é o momento da sondagem onde se seleccionam os elementos a partir dos quais se vão recolher os dados necessários para o estudo. Se seguida, contextualiza-se o problema em estudo de acordo com a teoria da amostragem.

Em grande parte das sondagens eleitorais utiliza-se amostragem estratificada. A amostragem estratificada é um processo de amostragem em que a população de dimensão N se encontra dividida em L estratos mutuamente exclusivos, sendo retirada uma amostra aleatória de n_i elementos para cada estrato i que possui N_i elementos. A dimensão da amostra total de n elementos é o somatório das sub-amostras, retiradas de cada estrato, $n = \sum_{i=1}^{L} n_i$. O objectivo ao estratificar uma população é reduzir a variabilidade dos estimadores e assim obter estimativas mais precisas, através da criação de grupos/estratos que originem grupos homogéneos a nível interno mas diferentes dos restantes grupos, fazendo com que desta forma a variabilidade da população seja fundamentalmente explicada pela variância entre os estratos, Reis e tal (2001).

Prosseguindo a ligação da teoria da amostragem com as sondagens eleitorais, tome-se a variável binária, Y, como sendo, por exemplo: "voto num determinado partido A", definida no espaço $S = \{0,1\}$ sendo 1 para quem vota no partido A e 0 para quem não vota no partido A. Variáveis deste tipo abundam na estimação em sondagens eleitorais. A proporção na população, p_i, de elementos no estrato i que vota no Partido A é dada por:

$$p_i = \frac{\sum_{j=1}^{N_i} Y_j}{N_i} \tag{2.1}$$

Esta proporção pode ser estimada por $\hat{p}_i$, sendo:

$$\hat{p}_i = \frac{\sum_{j=1}^{n_i} y_j}{n_i} \tag{2.2}$$

A estimativa a nível da população fornecida pela amostragem estratificada, $\hat{p}_{est}$, é dada por:

$$\hat{p}_{est} = \sum_{i=1}^{L} \frac{N_i \hat{p}_i}{N} \tag{2.3}$$

Segundo Cochran (1977), a variância de p_{est} é dada por:

$$V(\hat{p}_{est}) = \frac{1}{N^2} \sum_{i=1}^{L} \frac{N_i^2 (N_i - n_i)}{N_i - 1} \frac{p_i q_i}{n_i} \tag{2.4}$$

Em que $q_i = 1 - p_i$.

Como, de acordo com Cochran (1977) em quase todas as aplicações o termo $1/N_i$ será negligenciável, temos que:

$$V(\hat{p}_{est}) = \frac{1}{N^2} \sum_{i=1}^{L} N_i^2 \frac{p_i q_i}{n_i} - N_i p_i q_i \tag{2.5}$$

No caso da amostragem estratificada ser proporcional pelos *L* estratos:

$$V(\hat{p}_{est}) = \frac{N - n}{N} \frac{1}{nN} \sum_{i=1}^{L} \frac{N_i^2 p_i q_i}{N_i - 1} \tag{2.6}$$

Quando se estão a estimar as equações (2.5) e (2.6) a partir de uma amostra, temos que substituir o termo desconhecido $p_i q_i$ por $n_i \hat{p}_i \hat{q}_i / (n_i - 1)$, em que $\hat{q}_i = 1 - \hat{p}_i$.

Embora seja muitas vezes aconselhado utilizar-se um esquema de amostragem estratificada, uma sub-representação de alguns estratos amostrais pode levar à dificuldade de obter estimativas precisas. Em Portugal, por exemplo, nas eleições para a Assembleia da República, elegem-se deputados a nível de distrito. Tome-se uma amostra de n elementos, recorrendo a um processo de amostragem estratificada aleatória proporcional pelos diferentes distritos. Estimar p_{est}com um limite máximo de erro igual a B e um grau de confiança de $(1 - \alpha) \times 100\%$, implica ter uma estimativa que verifique:

$$P(|\hat{p}_{est} - p|) \leq B) = 1 - \alpha, \quad 0 < \alpha < 1 \tag{2.7}$$

Logo:

$$P(\hat{p}_{est} - B \leq p \leq \hat{p}_{est} + B) = 1 - \alpha \tag{2.8}$$

Por (2.6) e, uma vez que os valores reais de $p_i q_i$ são desconhecidos, temos que:

$$V(\hat{p}_{est}) = \frac{N-n}{N}\frac{1}{nN}\sum_{i=1}^{L}\frac{N_i^2}{N_i-1}\frac{n_i p_i q_i}{n_i-1} \tag{2.9}$$

Uma vez que segundo Reis et al (2001), pelo Teorema do Limite Central, podemos assumir que $\hat{p}_{est}$ possui uma distribuição normal temos que:

$$z\sqrt{V(\hat{p}_{est})} = B \tag{2.10}$$

Onde z é o valor da abcissa da curva normal que corta uma área de $(1-\alpha)$ nas caudas, como representado na Figura 1. Logo:

$$z\sqrt{\frac{N-n}{N}\frac{1}{nN}\sum_{i=1}^{L}\frac{N_i^2}{N_i-1}\frac{n_i p_i q_i}{n_i-1}} = B \tag{2.11}$$

Fixando uma confiança de 95,5% (o que implica $z = 2$), para o caso de $\hat{p}_i = 0{,}5, \forall i \in L$ (caso que implica $\hat{q}_i = 0{,}5$ e que define o máximo do produto $\hat{p}_i\hat{q}_i$, como se pode ver pela Figura 2) e considerando os valores na Tabela 4, para o caso específico de uma amostra de 997 elementos, obtida de forma proporcional por distrito, a nível de Portugal Continental, através do processo de amostragem aleatório estratificado temos que $B = 0{,}03196$.

Logo, temos o seguinte intervalo de confiança para $\hat{p}_{est}$:

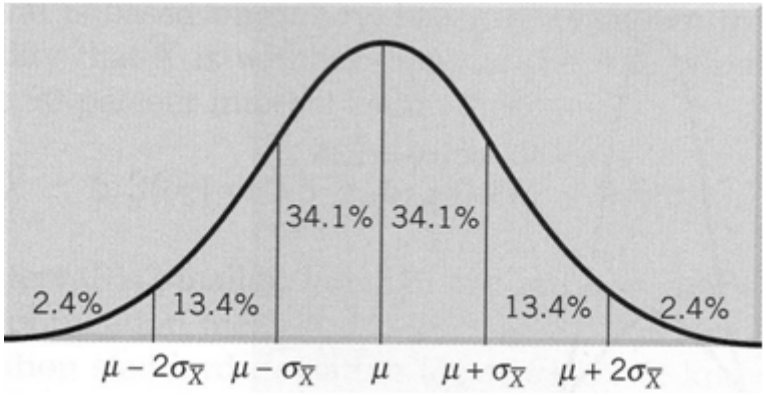
$$]\hat{p}_{est} - 0{,}03196;\ \hat{p}_{est} + 0{,}03196[\tag{2.12}$$

O que se traduz num erro amostral de aproximadamente 3,20%,

Tabela 1 – Distribuição da população e da amostra a nível continental.

Distrito	**Tamanho População (N_i)**	**Tamanho da amostra (n_i)**
Aveiro	592257	72
Beja	139437	17
Braga	670738	81
Bragança	129102	18
Castelo Branco	181128	23
Coimbra	379271	45
Évora	149533	17
Faro	337486	40
Guarda	155630	20
Leiria	386851	44
Lisboa	1820473	214
Portalegre	110067	13
Porto	1467249	172
Santarém	389442	44
Setúbal	669082	78
Viana do Castelo	212534	31
Vila Real	189800	26
Viseu	331329	42
Total	8311409	997

Figura 1 – Distribuição normal – Distribuição normal e relação existente entre z e o valor da abcissa. Na figura são visíveis os casos para $\mathbf{z = 1}$, caso da estimativa compreendida no intervalo $]\boldsymbol{\mu - \sigma_{\bar{X}}, \mu + \sigma_{\bar{X}}}[$ e para $\mathbf{z = 2}$, caso da estimativa compreendida no intervalo $]\boldsymbol{\mu - 2\sigma_{\bar{X}}, \mu + 2\sigma_{\bar{X}}}[$. Fonte: Aaker (2006)



Tome-se um exemplo em que a amostra de uma sondagem eleitoral proporcional tem $n = 997$ elementos, sendo obtida de forma aleatória estratificada de forma proporcional. Nesse caso, essa amostra representa, para o distrito de Portalegre, uma subamostra, obtida pelo método de amostragem aleatório simples, de $n = 13$ elementos. Considerando o tamanho desta subamostra, fixando um a confiança de 95,5% (o que implica $z = 2$) e considerando $\hat{p} = 0{,}5$ temos, por Cochran (1977):

$$B = z\sqrt{\frac{\hat{p}\hat{q}}{n-1}\frac{N-n}{N}} <=> B = 2 \times \sqrt{\frac{0{,}5 \times 0{,}5}{12} \times \frac{110067-1}{110067}} <=> B = 0{,}2887$$

O que se traduz num erro amostral de 28,87% para um intervalo de confiança a 95,5% e num intervalo de confiança:

$$]\hat{p} - 0{,}2887;\ \hat{p} + 0{,}2887[\qquad (2.13)$$

Qualquer inferência para o distrito em questão utilizando a subamostra tal como referida deve ser realizada com cautela, devido aos elevados erros amostrais envolvidos.

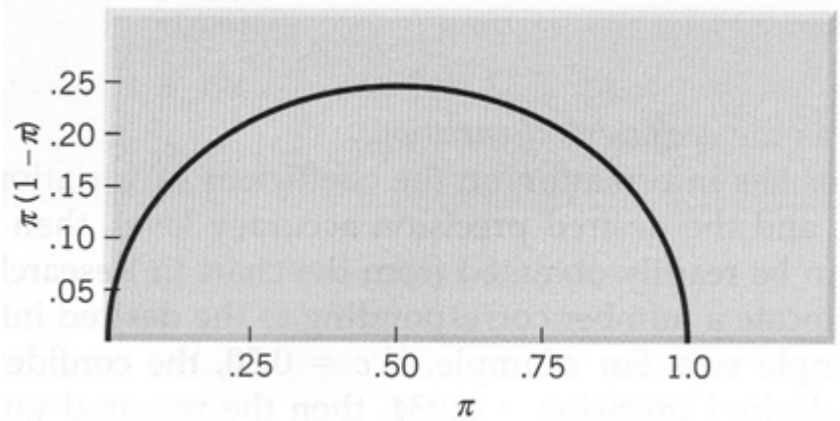
Uma amostra que forneça uma boa estimativa a nível nacional não é necessariamente uma amostra que fornece boas estimativas a nível mais desagregado, por exemplo a nível de distrito. Pela análise da Tabela 5, que apresenta os erros amostrais da amostra global e da subamostra de cada distrito, pode-se verificar que as subamostras a nível de distrito possuem erros que vão desde os 0,0685 (Lisboa) até os 0,2887 (Portalegre), o que mostra que as estimativas tendo por base estas subamostras poderão não ser muito precisas.

Actualmente, e cada vez mais, se procura diminuir os custos mantendo, ou mesmo aumentando, o benefício obtido. Se, através da aplicação de técnicas estatísticas, conseguirmos melhorar as estimativas de uma sondagem a um nível micro (município, distrito, região, etc.), podemos, no limite evitar a realização de uma nova sondagem, com o respectivo ganho a nível de custos e utilização de recursos. Este é um dos motivos pelos quais se poderá recorrer a métodos auxiliares de modo a melhorar as estimativas, obtidas por uma sondagem eleitoral (ou com outro propósito), a nível micro.

Tabela 2 – Erros amostrais por distrito – Erros amostrais da amostra obtida pelo processo de amostragem aleatório estratificado por distrito.

Distrito	Erro
Aveiro	0,1187
Beja	0,2500
Braga	0,1118
Bragança	0,2425
Castelo Branco	0,2132
Coimbra	0,1507
Évora	0,2500
Faro	0,1601
Guarda	0,2294
Leiria	0,1525
Lisboa	0,0685
Portalegre	0,2887
Porto	0,0765
Santarém	0,1525
Setúbal	0,1140
Viana do Castelo	0,1826
Vila Real	0,2000
Viseu	0,1562
Amostra Total	0,0319

Figura 2 – Relação entre p e pq – Gráfico da relação entre p e pq, na figura π e π(1-π), respectivamente. Fonte: Aaker (2006).



2.2 – Estimação em Pequenos Domínios

Os métodos de estimação em pequenos domínios permitem contornar os problemas associados à dificuldade em divulgar informação relativa a domínios de interesse para os quais não foram recolhidos dados que permitam obter estimativas fiáveis dos parâmetros que se pretendem estimar. Estes métodos são úteis quando a estimação directa falha devido a amostragens imprecisas ou inadequadas.

A aplicação destes métodos é hoje uma realidade em vários países e em diferentes áreas de pesquisa, desde a utilização para a obtenção de um Índice de Produção Industrial (IPI) para a comunidade autónoma Catalã pelo IDESCAT-UPF (Satorra e Ventura (2006)), passando pela sua aplicação em índices agrícolas na Turquia e até no Inquérito de Despesas Médicas nos Estados Unidos. São inúmeros os exemplos e possibilidades de aplicação. Uma das aplicações mais frequentes da estimação em pequenos domínios, já posta em prática em vários países do mundo, é a estimação das taxas de desemprego, realizada no âmbito dos diversos Inquéritos ao Emprego (Labour Force Surveys).

Para Rao, o termo “small area”, traduzido correntemente como “pequeno domínio” é comummente usado para definir uma área geograficamente pequena, tal como uma determinada região ou um concelho. Pode também ser usado para definir um outro tipo de “pequeno domínio”, isto é, uma subpopulação de doentes com determinada patologia ou um grupo etário específico.

No âmbito da estimação em pequenos domínios têm sido propostos diferentes métodos, sendo possível classificar esses métodos de várias formas. Uma das mais utilizadas é a proposta por Rao (Rao, 2003) que sugere a seguinte classificação dos estimadores: directos, indirectos, sintéticos e combinados. Segundo Satorra e Ventura (2006), os métodos para estimação em pequenos domínios incluem estimadores directos, estimadores sintéticos e outros estimadores indirectos.

Os estimadores directos utilizam apenas dados provenientes do pequeno domínio estudado. Usualmente estes estimadores não são enviesados, mas possuem uma elevada variância. Os estimadores indirectos e compostos são mais precisos, uma vez que

utilizam dados provenientes de variáveis relacionadas ou de áreas vizinhas. Estimadores indirectos são obtidos usando estimadores relacionados com todo o domínio e que são não enviesados. Os estimadores compósitos são combinações lineares de estimadores directos e indirectos.

Podemos descrever e sintetizar as diferenças entre os diversos tipos de estimadores de pequenos domínios da seguinte forma (Coelho, (2008)):

i) **Estimadores Directos**: estimadores que apenas utilizam informação das variáveis de interesse pertencentes ao pequeno domínio em estudo. Os estimadores directos são aqueles que são obtidos através da aplicação de pesos às unidades amostrais seleccionadas em cada área. Contudo, como muitas sondagens são desenhadas de modo a obter estimativas representativas apenas a um nível geográfico mais elevado, os tamanhos de amostra, a um nível de desagregação mais baixo, são, frequentemente, pequenos para produzirem estimativas directas precisas.

ii) **Estimadores Indirectos**: são estimadores que utilizam informação das variáveis de interesse e de variáveis auxiliares fora do pequeno domínio em estudo (ou do âmbito temporal em causa). Este tipo de estimadores pode-se subdividir em:

- **Estimadores directos modificados**: são estimadores que mantêm propriedades estatísticas interessantes sob amostragem repetida.
- **Estimadores sintéticos**: são estimadores cujas propriedades dependem dos pressupostos baseados dos modelos definidos.
- **Estimadores combinados**: resultam de uma combinação linear entre um estimador directo e um estimador sintético. Podem-se designar também por estimadores compostos ou compósitos. Estes estimadores possuem uma combinação entre um termo design based (associado aos estimadores directos e directos modificados) e um termo model based (associado aos estimadores sintéticos). Como exemplo destes estimadores temos: o EBLUP, o estimador de Fay-Herriot (modelo a nível área) e o estimador de Battese-Harter-Fuller (modelo a nível unidade).

Os estimadores directos são aqueles que se baseiam apenas nos dados específicos do domínio de interesse e exclusivamente para o período temporal de interesse. Os estimadores indirectos baseiam-se em dados fora do domínio ou período temporal de interesse e têm a vantagem de produzirem estimativas geralmente não enviesadas, embora falhem em pequenas amostras, para as quais a variância do estimador é elevada.

Os estimadores indirectos podem também ser designados por sintéticos, model based ou estimadores de pequenos domínios, Rao (2003). Podem identificar-se três tipos diferentes de estimadores indirectos: i) Estimadores indirectos de domínio – utiliza-se apenas informação relativa a outros domínios de interesse, mas não de outros períodos temporais; ii) Estimadores indirectos de tempo - utiliza-se apenas informação relativa a outros períodos temporais, mas não de outros domínios de interesse; iii) Estimadores indirectos de domínio e de tempo – utiliza-se informação quer de outros domínios, quer de outros períodos temporais. A existência de informação auxiliar e a ligação a modelos adequados é fundamental para a formação de estimadores indirectos.

Os estimadores indirectos baseados em modelos de pequenos domínios designam-se por estimadores model-based. Por sua vez, os modelos de pequenos domínios podem dividir-se em: (i) modelos agregados ou de área, quando relacionam estimadores de pequenos domínios directos com variáveis agregadas relacionadas com o domínio geográfico de interesse; (ii) modelos elementares, quando se relacionam variáveis ao nível das unidades elementares da análise.

A estimação model-based para pequenos domínios possui várias vantagens, entre as quais o facto de serem deduzidos a partir de um modelo explícito, que é previamente seleccionado e devidamente validado a partir dos dados provenientes da amostra. Dependendo da natureza das variáveis de resposta e da complexidade da estrutura espacial e temporal dos dados, pode proceder-se ao teste de vários modelos. Os tipos de estimadores indirectos model-based mais utilizados são:

- Estimadores EBLUP (Empirical Best Linear Unbiased Prediction)
- Estimadores de Bayes empíricos paramétricos (EB)
- Estimadores de Bayes hierárquicos paramétricos (HB)

Os estimadores sintéticos são obtidos pelo ajustamento de um modelo aos dados disponíveis (aproximação model-based), sendo frequentemente um modelo de regressão. Funcionam geralmente bem, mesmo em pequenas amostras pois baseiam-se nas regressões subjacentes entre as variáveis do modelo. No entanto, o sucesso destes modelos depende da relação entre as variáveis do modelo: quanto maior for a correlação entre as variáveis (da população a estimar e as outras variáveis do modelo), melhor será a qualidade da estimativa.

Por último, os estimadores combinados são estimadores que resultam de uma combinação linear entre um estimador directo e um estimador sintético. Estes estimadores podem também designar-se por estimadores compostos ou compósitos, pois incorporam uma combinação entre um termo design-based (associado aos estimadores directos e directos modificados) e um termo model-based (associado aos estimadores sintéticos). São exemplos de estimadores combinados (Rao, (2003)) os estimadores EBLUP (Empirical Best Linear Unbiased Prediction), o GREG, o estimador de Fay-Herriot (modelo de nível área) e o estimador de Battese-Harter-Fuller (modelo de nível unidade)

Uma das vantagens da utilização de estimadores compósitos (tais como o GREG) é a realização de ajustamentos para permitir estimar as diferenças entre a amostra e a população para as diferentes áreas. No caso particular do EBLUP, este estimador reduz o "ruído" existente através da multiplicação dos resíduos por um factor γ. De facto o estimador EBLUP estima u_d, (os resíduos desconhecidos ao nível da população) com base em ε_d (os resíduos obtidos pela fixação do modelo).

Existem dois tipos de estimadores EBLUP, classificados quanto ao estabelecimento de uma relação de ligação ao nível das unidades amostrais no que respeita à informação auxiliar a considerar:

- De nível individual (EBLUP A): obriga a existência de uma função de ligação entre a informação directa e a informação auxiliar ao nível dos indivíduos;
- De nível área (EBLUP B): a informação auxiliar é agregada ao nível da área para a qual se pretendem calcular as estimativas.

Neste trabalho aplica-se o estimador EBLUP de nível área, sendo Y a variável a estimar e X a variável contendo informação auxiliar, temos:

$$\hat{\bar{Y}}_i = \mu_i + \varepsilon_i, \text{ com } \varepsilon_i \sim N(0; \hat{\sigma}_i^2) \tag{2.14}$$

Sendo a verdadeira média para a área i dada por:

$$\mu_i = \beta_0 + \beta_1 \bar{X}_i + Zu_i \text{ com } u_i \sim N(0; \sigma_u^2) \tag{2.15}$$

- μ_i representa os efeitos aleatórios e Z reflecte a estrutura dos efeitos aleatórios
- A estimação do modelo pode ser feita por Máxima Verosimilhança (ML) Máxima Verosimilhança Restrita (REML)

Assume-se $\hat{\bar{Y}}_i \sim N(\mu; V)$

As estimativas dos pequenos domínios são dadas por:

$$\hat{\bar{Y}}_i = \hat{\beta} X_i + \hat{u}_i \tag{2.16}$$

A variabilidade do estimador é dada pelo erro quadrático médio (MSE), calculado da seguinte forma:

$$\mathrm{MSE}\left[\hat{\bar{Y}}_i\right] \approx G1 + G2 + 2.G3 \tag{2.17}$$

Em que:

- G1 representa a incerteza da estimativa
- G2 representa a incerteza da estimação dos β
- G3 representa a incerteza da estimação de $\hat{\sigma}_u^2$

É frequente que somente depois de uma pesquisa/sondagem ter sido projectada e realizada, sejam definidos os domínios, ou áreas de interesse. Nestes casos, o dilema do estatístico é produzir estimativas fiáveis sem ter tido a hipótese de recolher os dados necessários. Uma solução razoável para melhorar as estimativas é o uso de dados provenientes de outras fontes.

A estimação em domínios deve seguir as seguintes fases (Coelho, P., 2008):

a) Identificação dos domínios (pequenos domínios);
b) Identificação das fontes relevantes de informação
   i. Dados administrativos
   ii. Dados censitários
   iii. Dados provenientes de inquéritos por amostragem relacionados
c) Escolha dos métodos de combinação da informação
   i. Estimação directa modificada
   ii. Estimação sintética
   iii. Estimação compósita (combinada)
      - Modelos de nível unidade
      - Modelos de nível área

No presente trabalho, será aplicado o estimador EBLUP, o que corresponde a seguir, de acordo com o esquema anterior, os passos a), b) e c) iii (Modelos de nível área).

2.3 - Regressão Multinível

A necessidade de incorporar ao mesmo tempo informação a nível individual e informação a nível do grupo a que os indivíduos de uma amostra pertencem levou muitos investigadores de várias áreas do saber à procura de novas técnicas estatísticas. Um dos aspectos mais desafiadores desta incorporação é a integração de informação micro e macro num único modelo (Leew et al, (2008)). Por exemplo, no caso das amostras provenientes da área da educação, os alunos encontram-se agrupados em turmas, turmas estas que se encontram agrupadas em escolas, que por sua vez se encontram agrupadas em distritos, e por ai adiante, numa sucessão hierárquica de agrupamentos progressivos.

Se o objectivo for a procura de técnicas que permitam inferir sobre a relação de uma variável dependente com base em variáveis preditoras ou independentes, podemos considerar que existem variáveis preditoras (independentes) para variáveis (dependentes) em todos estes níveis. O desafio reside em combinar todos estes preditores numa análise estatística apropriada, mais especificamente numa análise de regressão "integrada". É por isso que a regressão se designa, neste caso, **regressão multinível**, como se verá de seguida. Vários autores abordam esta metodologia multinível. Segue-se uma pequena síntese da revisão de literatura existente sobre este método.

Segundo Kreft et al (1998), dados com estruturas hierárquicas são muito comuns nas ciências sociais e comportamentais Uma hierarquia consiste em ter observações de baixo nível agrupadas em níveis mais elevados. Temos como exemplo empregados agrupados em firmas e eleitores agrupados em distritos. Os modelos multinível são desenvolvidos para analisar dados com estruturas hierárquicas.

Para Snijders et al (1999) a análise multinível é uma metodologia para a análise de dados com padrões de variabilidade complexos, com um foco de fontes localizadas de variabilidade. Geralmente, na análise destes casos é esclarecedor ter em consideração a variabilidade associada a cada nível de localização da fonte. A análise multinível é uma

abordagem ao estudo deste tipo de dados, que incluí as técnicas estatísticas e a sua metodologia de aplicação.

No caso das eleições para a Assembleia da República, citado na secção 2.1, está-se presente uma estrutura hierárquica, pois os eleitores encontram-se agrupados em concelhos, que por sua vez se encontram agrupados em distritos, que juntos formam a totalidade do Continente Português.

De acordo com Leeuw et al (2008), este tipo de problemas já foi abordado previamente, quer através da agregação a nível de grupo das variáveis a nível individual, (por exemplo, tomando a média de uma variável a nível individual como valor para o grupo) quer através da desagregação a nível individual das variáveis a nível de grupo (por exemplo, considerando o valor a nível de grupo como valor para todos os indivíduos pertencentes ao grupo). Era contudo claro que estas duas estratégias eram desagradáveis *ad hoc* e que podiam introduzir sérios enviesamentos. A integração dos resultados destas análises, usando variáveis a nível de grupo em regressões a nível individual, foi designada como análise contextual, ou regressão ecológica ("*ecological regression*").

A ênfase foi então colocada na análise de regressão e em dados com dois níveis de hierarquia, por exemplo alunos e escolas. A definição de um modelo de regressão para cada escola de modo separado não era satisfatório, porque frequentemente as amostras dentro das escolas eram pequenas e os coeficientes de regressão instáveis. Estas análises separadas ignoravam também o facto de que todas as escolas eram parte do mesmo sistema escolar e, consequentemente, era natural supor que os coeficientes de regressão seriam semelhantes. Esta similaridade deveria ser usada, de alguma forma, para melhorar a estabilidade dos coeficientes de regressão através do método que ficou conhecido por "borrowing strength".

Em estudos de larga escala com milhares de escolas e longas listas de coeficientes de regressão, estes não podiam ser considerados uma redução de dados suficiente para ser de alguma forma útil. Por outro lado, exigir que os coeficientes de regressão fossem semelhantes em todas as escolas era geralmente visto como uma condição muito

restritiva, uma vez que existe um número muito grande de razões pelas quais as regressões dentro das escolas possam diferir. Em algumas escolas, as pontuações nos testes eram relativamente importantes, enquanto em outras o estatuto socio-económico era um preditor muito mais importante. As escolas claramente diferiam quer em termos médios quer na variância do sucesso escolar. Claro que exigir que os coeficientes de regressão fossem semelhantes para todas as escolas providenciava uma redução de dados considerável e uma variância amostral pequena, mas o sentimento era de que os coeficientes de regressão obtidos eram enviesados e desprovidos de significado.

Assim, era necessária alguma forma de análise intermédia, que não resultasse num único conjunto de coeficientes de regressão, mas que também não estabelecesse um modelo separado para cada escola. Isto levou naturalmente à ideia de modelos com coeficientes variáveis nas diferentes escolas, mas deixou em aberto o problema de combinar preditores de diferentes níveis numa única técnica. No início dos anos 80, surge a ideia de utilizar, num primeiro nível, os coeficientes dos diferentes modelos de regressão dentro das escolas como variáveis dependentes e, numa segunda fase de regressão, os preditores a nível de escola. Mas nesta segunda fase, o modelo standard de regressão que assumia observações independentes não podia ser utilizado, sobretudo, porque resultava em estimativas dos coeficientes de regressão ineficientes e de estimativas enviesadas dos seus erros padrão.

Para Leeuw et al (2008), foi em meados dos anos 80 que se tornou claro que os modelos, procurados pelos investigadores educacionais já existiam havia algum tempo, e eram aplicados em outras áreas da estatística. Contudo, eram utilizados sob nomes diferentes, ou de uma forma ligeiramente diferente. Estes modelos eram conhecidos como modelos lineares mistos ou, num contexto Bayesiano, como modelos hierárquicos lineares. A constatação de que os problemas da análise contextual podiam ser embutidos neste quadro de modelos clássicos lineares deu origem àquilo que hoje se denomina de análise multinível.

Segundo Gelman et al (2007), A modelação multinível pode ser vista de duas formas equivalentes:

(i) Uma generalização da regressão linear, onde a coordenada na origem e os declives podem variar por grupos. Por exemplo, começando por um modelo de regressão com uma variável independente, x (um preditor),
$y_i = \alpha + \beta x_i + \varepsilon_i$, podemos generalizá-lo para o modelo de variação da coordenada na origem, $y_i = \alpha_{j[i]} + \beta x_i + \varepsilon_i$, e para o modelo de variação da coordenada na origem e do declive, $y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \varepsilon_i$, em que $j[i]$, denota o grupo do individuo i.

(ii) Equivalentemente, podemos ver a modelação multinível como uma regressão que inclui uma variável de input categórica que representa o grupo a que determinada observação pertence. Desta perspectiva, o indicador do grupo é um factor com J níveis, correspondentes a J preditores no modelo de regressão (ou $2J$ se interagirem com um preditor x na variação da coordenada na origem ou $3J$ se interagirem com dois preditores x_1, x_2, etc.)

Em qualquer dos casos, $J - 1$ preditores lineares são adicionados ao modelo (ou posto de outra forma, o termo constante nos modelos de regressão é substituído por J termos separados de coordenada na origem. O passo crucial na modelação multinível é que estes J coeficientes são eles próprios também modelados (de modo mais simples, atribuindo uma distribuição comum para os J parâmetros α_j, ou, mais geralmente, um modelo de regressão para os α_j dados preditores a nível de grupo). O modelo a nível de grupo é estimado simultaneamente com o modelo de regressão a nível individual.

Uma vez que a aplicação prática deste trabalho é a melhoria das previsões relativas à intenção de voto num dado partido, e que esta se trata de uma variável categórica, estamos interessados em aprofundar e aplicar o modelo multinível de regressão logística. Como tal, na próxima secção faz-se a apresentação formal deste modelo. Na secção 2.3.2 aborda-se a metodologia aplicada por Gelman (1997) e (2007), que consiste na aplicação de um modelo hierárquico (multinível) de regressão logística a uma variável categórica, seguido de uma pós-estratificação das previsões obtidas.

2.3.1 – Formalização da Regressão Logística Multinível

Na regressão logística, o objectivo é obter a predição de valores tomados por uma variável categórica, frequentemente binária, a partir de uma série de variáveis explicativas contínuas e/ou binárias. Segundo Gelman et al (2007), os modelos multinível são aplicados à regressão logística e a outros modelos lineares generalizados da mesma forma que à regressão linear: os coeficientes são agrupados em grupos e uma distribuição de probabilidade é atribuída a cada grupo. Ou de modo equivalente, termos de erro são adicionados ao modelo correspondendo a diferentes fontes de variação dos dados.

Reflectindo na utilidade da análise multinível e na importância das variáveis categóricas em muitas áreas da investigação, a generalização dos modelos multinível para variáveis categóricas tem sido, para Leeuw et al (2008), uma área activa na investigação estatística. Tal não é de estranhar, uma vez que as estruturas hierárquicas existem em quase todas as áreas do saber, como foi referido no início desta secção.

Para o caso de variáveis dicotómicas, foram desenvolvidas várias abordagens adoptando quer um modelo de regressão logística, quer um modelo de regressão *Probit*, bem como vários métodos para incorporar e estimar a influência dos efeitos aleatórios.

Nesta secção utilizaremos preferencialmente a formalização de Snijders et al (1999), em que a estrutura básica da regressão logística a dois níveis é uma colecção de N grupos (unidades no nível dois - área), com, em cada grupo, uma amostra aleatória de n_j unidades do nível um (indivíduos). A variável objectivo (dependente) é dicotómica e denotada por Y_{ij} para a unidade de nível um i no grupo j. O tamanho total da amostra é:

$$M = \sum_j n_j \tag{2.18}$$

Se não tivermos em conta variáveis explicativas (preditores), a probabilidade de sucesso é constante em cada grupo, sendo denotada por p_j. Num modelo de coeficientes aleatórios, os grupos são considerados provenientes de uma população de grupos e a

probabilidade de sucesso nos grupos, P_j, são consideradas como variáveis aleatórias definidas nesta população. O resultado dicotómico pode ser representado como a soma desta probabilidade com um resíduo, R_{ij}, logo temos:

$$Y_{ij} = P_j + R_{ij} \tag{2.19}$$

O resultado para o indivíduo i no grupo j, que poderá ser 0 ou 1, é expresso como a soma das probabilidades (proporção média de sucesso) no grupo mais um resíduo dependente do nível individual. Este resíduo possui média zero e a peculiaridade de poder assumir apenas os valores "$-P_j$" e "$1 - P_j$", uma vez que deve ser 0 ou 1. Outra particularidade é o facto de, dado o valor da probabilidade P_j, a variância do resíduo é:

$$V(R_{ij}) = P_j(1 - P_j) \tag{2.20}$$

Como a variável se encontra codificada como 0 ou 1, a média do grupo, $\hat{Y}_{.j}$ é dada por:

$$\hat{Y}_{.j} = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij} \tag{2.21}$$

e é a proporção de sucesso no grupo j. Esta é uma estimativa para a probabilidade dependente do grupo P_j. De modo similar, a média global, $\hat{P}_{.}$, é dada por:

$$\hat{P}_{.} = \hat{Y}_{..} = \frac{1}{M}\sum_{j=1}^{N}\sum_{i=1}^{n_j} Y_{ij} \tag{2.22}$$

e é a proporção global de sucesso. A variância entre as probabilidades dependentes dos grupos, isto é, o valor populacional $V(P_j)$, pode ser estimado por $\hat{\tau}^2$, que é obtido através de:

$$\hat{\tau}^2 = S^2_{entre} - \frac{S^2_{dentro}}{\tilde{n}} \tag{2.23}$$

em que S^2_{entre} é a variância observada entre grupos, S^2_{dentro} a variância dentro dos grupos, e em que:

$$\tilde{n} = \frac{1}{N-1}\left\{M - \frac{\sum_j n_j^2}{M}\right\} = \bar{n} - \frac{s^2(n_j)}{N\bar{n}} \tag{2.24}$$

$$\bar{n} = \frac{M}{N} \tag{2.25}$$

Para variáveis resultado dicotómicas, a variância observada entre grupos, S^2_{entre}, é dada por:

$$S^2_{entre} = \frac{\hat{P}_{.}(1-\hat{P}_{.})}{\tilde{n}(N-1)} X^2 \tag{2.26}$$

em que:

$$X^2 = \sum_{j=1}^{N} n_j \frac{(\bar{Y}_{.j} - \hat{P}_{.})^2}{\hat{P}_{.}(1-\hat{P}_{.})} \tag{2.27}$$

A variância dentro dos grupos, S^2_{dentro}, é dada por:

$$S^2_{dentro} = \frac{1}{M-N} \sum_{j=1}^{N} n_j \bar{Y}_{.j}(1-\bar{Y}_{.j}) \tag{2.28}$$

Snijders et al (1999), consideram ainda que pode ser relevante incluir variáveis explicativas no modelo. Quando tal sucede, um problema que surge é que as probabilidades estão restritas ao intervalo de 0 a 1, e um efeito linear para uma variável explicativa pode levar o resultado para fora deste intervalo. Para ultrapassar este problema, em vez da probabilidade de algum evento, pode-se considerar o rácio entre a probabilidade de sucesso e a probabilidade de insucesso. Quando a probabilidade de sucesso é p, este rácio é $p/(1-p)$. Este valor, ao contrário das probabilidades, pode assumir qualquer valor de 0 a infinito e pode ser considerado uma escala de rácio. O logaritmo transforma uma escala multiplicativa numa escala aditiva e transforma o conjunto dos números reais positivos na recta real, desta forma:

$$logit(p) = \ln\left(\frac{p}{1-p}\right) \tag{2.29}$$

A função *logit* é uma função definida para números de 0 a 1 e o seu contra-domínio é IR. O modelo de regressão logística é um modelo em que $logit(p)$ é uma função linear das variáveis explicativas. O termo geral para este tipo de funções é função *link*, uma vez que liga as probabilidades às variáveis explicativas. A função *probit* é também frequentemente usada como uma função *link* para variáveis dicotómicas. A escolha da função *link* deve, segundo Snijders et al (1999), ser guiada por um ajuste empírico do modelo, facilidade de interpretação e conveniência (por exemplo, existência de software

para a implementação de modelos). Para a selecção da função *link* ser relevante, Leeuw et al (Donald Hedeker - 2008), referem que é necessária uma grande quantidade de dados. Uma vez que estas funções frequentemente fornecem ajustes e conclusões similares, a escolha da função poderá ser feita primariamente por motivos de facilidade de interpretação.

Para Snijders et al (1999), o "modelo vazio" a dois níveis para uma variável objectivo dicotómica refere-se à população de grupos (unidades do nível dois) e especifica a distribuição de probabilidade para as probabilidades dependentes dos grupos, P_j, sem ter em consideração qualquer variável exploratória. Vários métodos e especificações desta distribuição foram sugeridos, mas os autores focam-se no método que especifica as probabilidades transformadas $f(P_j)$, como tendo uma distribuição normal. Isto é expresso, por meio de uma função *link* geral $f(p)$, pela fórmula:

$$f(P_j) = \gamma_0 + U_{0j} \tag{2.30}$$

onde γ_0 é a média populacional das probabilidades transformadas e U_{0j} o desvio aleatório desta média para o grupo j. Se $f(p)$ é a função $logit$, então $f(P_j)$ é apenas o logaritmo do rácio de probabilidades para o grupo j. Desta forma, para a função $logit$, o logaritmo do rácio de probabilidades tem uma distribuição normal na população dos grupos, que é expressa por:

$$logit(P_j) = \gamma_0 + U_{0j} \tag{2.31}$$

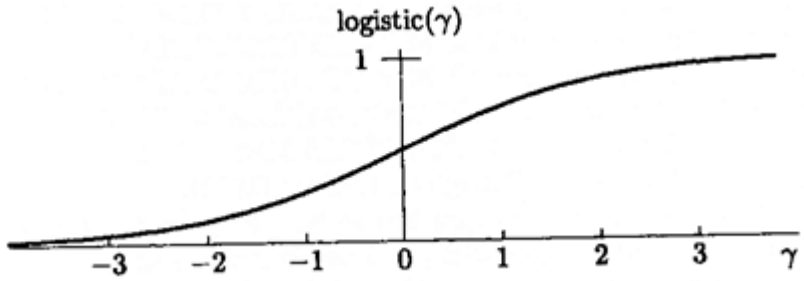
Os desvios U_{0j} assumem-se como variáveis aleatórias independentes com uma distribuição normal de média 0 e variância τ_0^2. Este modelo não inclui um parâmetro separado para variância de nível um porque a variância residual deste nível, para o caso de variáveis objectivo dicotómicas, segue directamente da probabilidade de sucesso.

Denote-se por π_0 a probabilidade correspondente à média γ_0, definida por: $f(\pi_0) = \gamma_0$. Para a função $logit$, isto significa que π_0 é a transformação logística de γ_0, definida por:

$$\pi_0 = logistic(\gamma_0) = \frac{\exp(\gamma_0)}{1 + \exp(\gamma_0)} \tag{2.32}$$

Note-se que a função logística e a função $logit$ são funções inversas. O gráfico seguinte mostra a forma da função logística.

Figura 3 – Função Logística, fonte: Snijders et al (1999)



O valor π_0 é aproximado (mas não igual) ao valor médio das probabilidades P_j na população dos grupos. Devido à natureza não linear da função *link*, não existe uma relação simples entre a variância destas probabilidades e a variância dos desvios U_{0j}.

Segundo Snijders et al (1999), A estimação dos parâmetros para estes modelos é mais complicada do que no caso dos modelos hierárquicos lineares, sendo que é inevitável a utilização de algum tipo de aproximação. Vários métodos de aproximação foram propostos, desde expansões de Taylor de primeira ou de segunda ordem para a função *link* até métodos computacionais intensivos relacionados com bootstrap e o método de amostragem de Gibbs.

2.3.2 – A Formalização de Gelman – Regressão Logística Multinível e Pós-Estratificação

Para este trabalho adoptou-se a formalização e a metodologia utilizada por Gelman (Gelman, (1997) e (2007)), pois assume-se como sendo completa e autocontida, surgindo como uma boa base para o desenvolvimento que se pretende com este estudo.

Na notação que se segue, considere-se uma partição da população em estudo em R variáveis categóricas, onde a r-ésima variável possui J_r níveis, para um total de $J = \Pi_{r=1} j_r$ categorias (células), que etiquetamos de $j = 1, \dots J$. Seja N_j, o número de indivíduos na população pertencentes à categoria j, e suponha-se que este valor é conhecido para todos os j. Seja y uma resposta binária de interesse, e seja π_j a média populacional em cada estrato j. Para cada j, seja n_j o número de indivíduos na categoria j na amostra. As R variáveis deverão incluir toda a informação usada para construir os pesos amostrais, bem como outras variáveis que poderão ser informativas relativamente a y.

Para Gelman et al (1997), podemos ajustar um modelo de regressão logística para a probabilidade π_j, de um "sim" na variável y, para respondentes na categoria j da seguinte forma:

$$logit(\pi_j) = X_j \beta \tag{2.33}$$

onde X é a matriz indicadora das variáveis, X_j é a j-ésima coluna de X e β é o vector dos parâmetros de regressão desconhecidos.

O modelo hierárquico (multinível) permite uma agregação parcial entre as células através da definição de um modelo de efeitos mistos. O modelo (2.33) pode então ser escrito na forma padrão de um modelo hierárquico de regressão logística como:

$$y_i \sim Binomial(p_i) \tag{2.34}$$

$$logit(p_i) = (X\beta)_i \tag{2.35}$$

$$\beta \sim N(0, \Sigma_\beta) \tag{2.36}$$

Onde Σ_{β}^{-1}é uma matriz diagonal com 0 para cada elemento de α, seguido de σ_m para cada elemento de γ_m, para cada m. Utiliza-se a notação p_i, para a probabilidade correspondente à unidade i, como distinção de π_j, a probabilidade agregada correspondente à categoria j. Escreve-se o vector β como $(\alpha, \gamma_1, \dots, \gamma_m)$, onde α é um subvector de coeficientes desagregados. Cada γ_m, para $m = 1, \dots, M$, é um subvector de coeficientes, γ_{km}, para o qual ajustamos um modelo hierárquico:

$$\gamma_{km} \overset{ind}{\sim} N(0, \sigma_m^2), k = 1, \dots, K_m \tag{2.37}$$

Dado que se incluiu um termo constante como um dos efeitos fixos, α, é razoável atribuir a cada γ_{km} uma média *a priori* de 0. Os parâmetros de variância, σ_m, são considerados com distribuições *a priori* independentes e não informativas:

$$\sigma_m^2 \sim inverse.gamma(0.001, 0.001), m = 1, .. M \tag{2.38}$$

Estas distribuições *a priori* independentes e não informativas permitem que os σ_m sejam estimados do conjunto de dados. Isto pode ser contrastado a dois extremos que correspondem à análise clássica. Ajustar σ_m como 0 corresponde a excluir um conjunto de variáveis, i.e., agregação completa o que se traduz num modelo de regressão normal sem considerar a estrutura hierárquica. Ajustar σ_m a ∞ corresponde a uma distribuição *a priori* não informativa para os parâmetros γ_{km}, ou seja não realizar agregação, o que se traduz na elaboração de um modelo para cada categoria j.

Para obter as quantidades de interesse, Gelman et al (2007) utilizam a seguinte estratégica de simulação Bayesiana:

- Realizar inferência Bayesiana para os coeficientes de regressão, β, e para os hiperparâmetros, σ_m, através do conjunto de dados.
- Para cada uma das J categorias na população, calcular $p_j = logit^{-1}(X\beta)_j$. Isto é feito para todas as categorias j, mesmo para as que não estão incluídas na amostra.
- Calcular inferências para as quantidades da população, somando $N_j p_j$'s dentro de cada estado.

O modelo é estimado através da simulação Bayesiana (Markov chain Monte Carlo methods), recorrendo ao WinBUGS. Estas simulações são depois utilizadas para calcular incertezas e erros padrão. Num contexto de regressão, a análise deve incluir como variáveis, tudo o que tenha um efeito importante na selecção da amostra ou nas não-respostas. Após a obtenção das estimativas de interesse, Gelman et al (2007) procedem, após a definição do modelo hierárquico e sua simulação, a uma pós-estratificação para a obtenção de um ponto preditor (por exemplo média ou mediana) para sumariar as simulações. Segundo Gelman et al (2007), o método de amostragem de Gibbs é o nome dado a uma família de algoritmos interactivos que são utilizados pelo BUGS ("Bayesian Inference Using Gibbs Samplig") e outros programas para o ajuste de modelos Bayesianos. A ideia base da amostragem de Gibbs é particionar o conjunto de parâmetros desconhecidos e proceder à sua estimação um de cada vez, ou um grupo de cada vez, mas em que cada parâmetro ou grupo de parâmetros é estimado de forma condicional a todos os restantes.

O algoritmo pode ser definido da seguinte forma:

1. *Definir um número, n_{corr}, de simulações paralelas a correr (tipicamente um número pequeno). Para cada uma destas correntes:*
    a. *Começar com um valor inicial para todos os parâmetros. Estes devem ser dispersos, tipicamente obtidos aleatoriamente*
    b. *Definir um número, n_{it}, de iterações (tipicamente um número grande). Para cada iteração actualiza-se os parâmetros, ou grupos de parâmetros, um de cada vez. Para cada parâmetro/grupo de parâmetros toma-se uma simulação aleatória dada o conjunto de dados e a estimativa actual de todos os outros parâmetros.*
2. *Avaliar a mistura de correntes aleatórias utilizando uma estatística de convergência, $\hat{R}$, factor de redução potencial de escala, para cada parâmetro. Este parâmetro é aproximadamente a raiz quadrada da variância da mistura de todas as correntes, dividida pela variância média dentro-correntes. Gelman et al (2007), sugere tomar $\hat{R} < 1.1$ para indicação de convergência.*
3. *Se a convergência não é atingida, repetir aumentando o número de iterações ou proceder a uma alteração do modelo.*

Para Gelman et al (2007), a parte chave deste algoritmo é o passo de actualização sequencial automática.

Existem geralmente muitas opções disponíveis para modelar uma estrutura de dados, e uma vez ajustado, com sucesso, um modelo é importante verificar de que modo é que este se ajusta aos dados e frequentemente compará-lo com outros modelos. Para tal, estes autores sugerem alguns métodos relativamente simples. A monitorização da qualidade de um modelo através da detecção de diferenças sistemáticas entre o modelo e os dados observados, utilizando para tal a capacidade preditiva das simulações realizadas, é uma das opções apresentadas. Para a comparação entre modelos, Gelman et al (2007), sugerem a utilização do "desvio", $D(\theta)$, definido como:

$$D(\theta) = -2\log\ \mathrm{p}(y|\theta) \tag{2.39}$$

em que $\log\ \mathrm{p}(y|\theta)$ é o logaritmo da probabilidade do conjunto de dados. Os autores sugerem ainda a utilização do parâmetro DIC (critério de informação do desvio). Este parâmetro é fornecido de forma automática pelo WinBUGS e é dado por:

$$DIC = D\left(\hat{\theta}\right) + 2p_D \tag{2.40}$$

Em que p_D é o número efectivo de parâmetros e $D\left(\hat{\theta}\right) = E_{\theta|y}[\theta]$. Quanto mais baixo for o valor do DIC melhor será o ajuste ao conjunto de dados e a capacidade preditiva do modelo.

Uma vez fixado o modelo e obtidas as previsões pode-se recorrer a uma pós-estratificação de modo a corrigir a amostra de eventuais desvios em relação à distribuição populacional. Para Gelman et al (1997), o objectivo da pós-estratificação é a correcção de diferenças entre a amostra e a população. Na notação que se segue, as letras minúsculas são utilizadas para variáveis que são observadas apenas na amostra e as letras maiúsculas para as variáveis que são observadas na amostra mas conhecidas na população. Supondo que temos uma matriz de indicadores de variáveis, X, cujas distribuições conjuntas na população são conhecidas, e uma variável y, cuja distribuição populacional estamos interessados em estimar. Assumindo que as variáveis são discretas e catalogando as categorias de X como células de pós-estratificação j, com N_j elementos na população e n_j elementos na amostra. Com esta notação, o total da população é $N = \sum_{j=1}^{J} N_j$ e o tamanho da amostra é $n = \sum_{j=1}^{J} n_j$. Implícito no modelo

de pós-estratificação está que os dados são recolhidos por um processo aleatório simples em cada um dos J estratos. O tamanho das amostras dos estratos criados é irrelevante. O caso da estratificação (de onde as amostras são retiradas pelo processo de amostragem da sondagem) não é mais do que um caso específico de pós-estratificação da forma como está formulada. Assume-se que os totais dos estratos N_j para cada categoria j são conhecidos. Estas categorias incluem todas as classificações cruzadas dos preditores X.

A média da população, θ, de qualquer resposta pode ser escrita como uma soma sobre os estratos definidos:

$$\theta = \frac{\sum_{j=1}^{J} N_j \theta_j}{\sum_{j=1}^{J} N_j} \tag{2.41}$$

Em que θ_j é a média do estrato j. Logo a estimativa pela pós-estratificação, $\hat{\theta}^{ps}$, é:

$$\hat{\theta}^{ps} = \frac{\sum_{j=1}^{J} N_j \hat{\theta}_j}{\sum_{j=1}^{J} N_j} \tag{2.42}$$

O caso mais simples, e o utilizado neste trabalho, é a pós-estratificação total, em que a estimação para cada estrato j é a média estimada do estrato $\hat{\theta}_j = \bar{y}_j$, logo:

$$\hat{\theta}^{ps} = \frac{\sum_{j=1}^{J} N_j \bar{y}_j}{\sum_{j=1}^{J} N_j} \tag{2.43}$$

Segundo Lax et al (2009), a regressão multinível e a pós-estratificação incorporam informação demográfica para melhorar as estimativas a nível de estado, enquanto permite a existência de diferenças não demográficas entre estados. Isto é, a opinião é modelada como uma função de efeitos demográficos e efeitos específicos dos estados. A estimação destes efeitos é melhorada através da utilização de um modelo multinível e as predições são feitas para cada respondente tipo em termos de características demográficas-geográficas com um peso obtido em relação aos dados populacionais. A desvantagem é a necessidade de dados demográficos detalhados acerca dos respondentes e dos estados, bem como uma metodologia complexa comparativamente com outros métodos (por exemplo, a agregação de sondagens).

Os métodos de regressão multinível e de pós-estratificação foram aplicados por Park et al (2004) que, para a estimação da intenção de voto nas eleições presenciais de 1988 nos EUA, utilizaram todos os respondentes de sete sondagens nacionais pré-eleitorais conduzidas pela CBS News/New York Times durante os nove dias que antecederam as referidas eleições. Neste trabalho, foram definidos alguns modelos hierárquicos e confrontados os seus resultados com os da sondagem eleitoral. Apesar de, a nível nacional todos os modelos apresentarem resultados similares, a redução no erro absoluto médio nos estados varia de 10,3% (CBS/NYT) a 2.1% (um dos modelos hierárquicos definidos pelos autores). Para os autores, a pós-estratificação é mais útil para estados em que a amostra de dados é pequena.

Ainda para Park et al (2004), o objectivo último da modelação probabilística e da inferência Bayesiana no contexto de uma sondagem é possibilitar a utilização de informação de pós-estratificação existente (por exemplo, através dos Censos), para proceder a um ajuste de uma amostra aleatória relativamente pequena.

Gelman et al (1997), procede também à definição de um modelo hierárquico e posterior pós-estratificação, utilizando também sondagens nacionais pré-eleitorais conduzidas pela CBS News/New York Times para prever os resultados eleitorais nas eleições presidenciais dos EUA em 1988 e 1992 de modo a obter estimativas das percentagens de apoio a Bush.

2.4 – Agregação de Sondagens

Um dos métodos mais comuns para estimar opiniões a nível de estado é a agregação de sondagens, Lax et al (2009). A principal vantagem deste método é a sua simplicidade e facilidade de aplicação. Após a combinação de um conjunto de sondagens a nível nacional, elaboradas segundo a mesma metodologia, calcula-se as percentagens de interesse desagregadas por estado. A única informação necessária é a opinião e o estado de residência do entrevistado. A ideia base deste método é a obtenção de estimativas mais precisas através da obtenção de amostras com maiores dimensões provenientes da agregação das sondagens. Contudo, para os autores, caso a opinião a sondar não seja estável ao longo do tempo, a aplicação deste método terá piores resultados do que qualquer sondagem isolada num período de tempo específico.

Existem ainda outros potenciais problemas. Este método requer um número elevado de sondagens a nível nacional para se conseguirem tamanhos de amostras suficientemente grandes em cada estado. Uma vez que tal é complicado, pelo menos num curto espaço temporal, este método é preferível quando a opinião é relativamente estável ao longo do tempo, podendo mesmo originar erros graves a agregação de sondagens nos casos em que há alteração súbita de opinião. Outro problema existente é que as sondagens representativas a nível nacional não o são necessariamente a nível de estado/distrito, podendo a acumulação de sondagens resultar numa amostra não representativa a nível de estado/distrito.

Este método foi inicialmente utilizado por Erikson (1993), através da agregação de 122 sondagens da CBS/NYT a nível nacional realizadas de 1976 a 1988 de modo a estimarem a opinião pública nos EUA. Tal foi possível porque as sondagens foram realizadas numa base contínua, mantendo as mesmas questões sobre identificação partidária e ideologia ao longo do tempo. Através da agregação das 122 sondagens foi possível obter uma amostra total de 167 460 elementos a nível nacional, que foi depois desagregada a nível de estado para a obtenção de estimativas a este grau.

Este método pode ser formalizado da seguinte forma: considere-se uma população com I estratos, para a qual sondamos J amostras independentes para todos os estratos ou somente para alguns deles. Seja n_{ij} o tamanho da amostra para o estrato i da sondagem j. Após a agregação das sondagens, o tamanho total da amostra para o estrato i é a soma das amostras de cada sondagem e é dada por:

$$n_i = \sum_{j=1}^{J} n_{ij} \tag{2.44}$$

Seja $\hat{p}_{ij}$ um estimador não enviesado de p_i obtido pela sondagem j, a estimativa da proporção de elementos do estrato i que votam no partido A após a agregação das sondagens é:

$$\hat{p}_i = \frac{n_{i1}\hat{p}_{i1} + n_{ij}\hat{p}_{i2} + \cdots + n_{ij}\hat{p}_{ij}}{n_i} = \frac{1}{n_i}\sum_{j=1}^{J} n_{ij}\hat{p}_{ij} \tag{2.45}$$

Como a amostra total da agregação das sondagens é dada por:

$$N = \sum_{i=1}^{I} n_i = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} \tag{2.46}$$

A estimativa da proporção de elementos na população que votam no partido A é dada por:

$$\hat{p} = \frac{n_i\hat{p}_i + n_i\hat{p}_i + \cdots + n_i\hat{p}_i}{N} = \frac{1}{N}\sum_{i=1}^{I} n_i\hat{p}_i \tag{2.47}$$

Uma aplicação deste método, com base em diferentes sondagens será apresentada no capítulo 4.

Capítulo 3 - Aplicação ao caso português: estimativas distritais e nacionais

Neste capítulo apresenta-se uma aplicação ao caso português das metodologias descritas anteriormente para estimar a proporção de votantes a nível distrital e nacional. Utilizou-se um conjunto de dados de uma sondagem eleitoral realizada a nível continental (regiões autónomas não incluídas) para aplicar os métodos previamente apresentados. Os métodos alternativos são aplicados segundo a ordem: Agregação de Sondagens (3.2), Regressão Multinível (3.3) e Estimador EBLUP (3.4). Para a aplicação das técnicas foram utilizados os seguintes programas computacionais: SPSS 15, R (R Development Core Team (2010)), Excel 2007 e WinBUGS (Lunn (2010).

3.1 - Descrição do conjunto de dados

Os dados utilizados neste trabalho são referentes a uma sondagem efectuada pelo **IPOM – Instituto de Pesquisa e Opinião de Mercado** para o **Jornal O Independente** sobre a intenção de voto nas eleições legislativas portuguesas de 2005, tendo sido recolhidos em Fevereiro do mesmo ano, via telefone, através de um questionário elaborado especificamente para o efeito (ver anexo I), tendo como dimensão amostral 997 eleitores a nível nacional. A sondagem em questão foi publicada no referido jornal tendo, como se veio a verificar após as eleições, alcançado resultados muito próximos dos resultados eleitorais referentes à população [2].

[2] Este facto foi confirmado empiricamente e encontra-se descrito, por exemplo, no blog do Prof. Pedro Magalhães, ex-responsável pelo CESOP (Centro de Sondagens da Universidade Católica), em: http://margensdeerro.blogspot.com/2005/03/contas-finais-mtodo-3.html.

O conjunto de dados era constituído inicialmente por 18 variáveis, mas para a aplicação das técnicas restringiu-se o conjunto de dados a um conjunto inferior de variáveis e procedeu-se a algumas recodificações das mesmas. Tal sucedeu devido à existência de variáveis desnecessárias à aplicação dos métodos que são utilizadas no questionário por diversas razões:

- Para a recolha de outra informação considerada relevante
- Para ir ganhando a confiança do entrevistado ao longo do questionário no sentido de diminuir a taxa de não respostas em questões mais sensíveis como a intenção de voto;

Estas alterações no conjunto de dados foram também necessárias devido à necessidade deste conjunto estar em certas forma padrão de codificação para a aplicação correcta dos métodos.

Desta forma, as variáveis e codificações consideradas foram as apresentadas de seguida e a amostra ficou reduzida a 719 elementos. De seguida apresentam-se as variáveis (questões) consideradas.

Variável 1 - Em qual dos seguintes partidos pensa votar no próximo domingo? (apresentada na Tabela 3)

Tabela 3 – Codificação da variável 1 - Em qual dos seguintes partidos pensa votar no próximo domingo?

Codificação	Voto
1	PSD
2	PS
3	CDS/PP
4	CDU
5	BE
6	Outros
7	Branco/Nulo
8	Indeciso
9	Não responde
10	Não vai votar

Para a aplicação dos métodos optou-se pela estimação da proporção de elementos que votam no PS e, após se retirar da análise as observações que não responderam (opção 9), não votariam (opção 10) ou ainda não tinham decidido a sua intenção de voto (opção 8), recodificou-se a variável voto da forma apresentada na Tabela 4.

Tabela 4 – Recodificação da variável 1 - Em qual dos seguintes partidos pensa votar no próximo domingo?

Codificação	Voto
0	Outra
1	PS

Variável 2 - Sexo com 2 categorias (0 - Masculino/1 - Feminino)

Variável 3 - Idade. Esta variável foi codificada de acordo com a Tabela 5.

Tabela 5 – Codificação da variável 3 - Idade

Codificação	Idade
1	De 18 a 24 anos
2	De 25 a 29 anos
3	De 30 a 39 anos
4	De 40 a 49 anos
5	De 50 a 59 anos
6	Com mais de 60 anos

Variável 4 - Distrito. Esta variável foi codificada de acordo com a Tabela 6.

Tabela 6 – Codificação da variável 4 - Distrito

Codificação	Distrito
1	Aveiro
2	Beja
3	Braga
4	Bragança
5	Castelo Branco
6	Coimbra
7	Évora
8	Faro
9	Guarda
10	Leiria
11	Lisboa
12	Portalegre
13	Porto
14	Santarém
15	Setúbal
16	Viana do Castelo
17	Vila Real
18	Viseu

Variável 5 - Exerce alguma actividade profissional? (0 – Não/1 - Sim)

Variável 6 - Formação Académica. Esta variável foi codificada de acordo com a Tabela 7.

Tabela 7 – Codificação da variável 6 – Formação Académica

Codificação	Formação Académica
1	Analfabeto
2	Sabe Ler/Primária
3	Até ao 9º ano
4	Até ao 12º ano
5	Bacharel
6	Licenciado
7	Pós-Graduado
8	Doutorado

Desta forma o conjunto final de dados, com 6 variáveis e 719 elementos, apresenta o aspecto descrito na Tabela 8.

Tabela 8 – Aspecto geral do conjunto de dados.

Voto	Sexo	Idade	Distrito	Exerce actividade profissional	Escolaridade
PS	Feminino	Entre 40 a 50 anos	Porto	Sim	Até ao 9º ano
PS	Masculino	Com mais de 60 anos	Porto	Não	Sabe Ler/Primária
…	…	…	…	…	…

A Tabela 9 traduz a distribuição, por distrito, da amostra a nível continental.

Tabela 9 – Distribuição, por distrito, da amostra da sondagem a nível de continente

Distrito	n	%
Aveiro	43	6
Beja	14	1,9
Braga	66	9,2
Bragança	10	1,4
Castelo Branco	16	2,2
Coimbra	26	3,6
Évora	12	1,7
Faro	27	3,8
Guarda	12	1,7
Leiria	39	5,4
Lisboa	154	21,4
Portalegre	8	1,1
Porto	136	18,9
Santarém	29	4
Setúbal	57	7,9
Viana do Castelo	24	3,3
Vila Real	17	2,4
Viseu	29	4

Os resultados, por distrito, obtidos nesta sondagem encontram-se descritos na Tabela 10, que apresenta as frequências relativas verificadas pelo PS a nível distrital.

Tabela 10 – Resultados da sondagem, por distrito

Distrito	Freq. Rel. PS	Outro
Aveiro	0,3256	0,6744
Beja	0,2143	0,7857
Braga	0,4697	0,5303
Bragança	0,4000	0,6000
Castelo Branco	0,8125	0,1875
Coimbra	0,5769	0,4231
Évora	0,6667	0,3333
Faro	0,4444	0,5556
Guarda	0,6667	0,3333
Leiria	0,2308	0,7692
Lisboa	0,4351	0,5649
Portalegre	0,8750	0,1250
Porto	0,4412	0,5588
Santarém	0,5862	0,4138
Setúbal	0,4737	0,5263
Viana do Castelo	0,5000	0,5000
Vila Real	0,4118	0,5882
Viseu	0,6207	0,3793
Previsão Continente	0,4655	0,5345
Erro Abs Médio	0,1221	

A previsão a nível de continente é o valor estimado pela amostra aleatória estratificada da sondagem.

3.2 - Aplicação do Método de Agregação de Sondagens

Nesta secção apresentam-se os resultados e conclusões obtidos através da aplicação de uma variação do método de agregação de sondagens, tendo-se, para tal, feito uso de duas sondagens auxiliares a nível distrital, para além da sondagem a nível do continente referida na secção 3.1.

Neste estudo, a metodologia de desagregação de sondagens, teve uma relativa variação face ao trabalho desenvolvido por Erikson et al (1993) no seu trabalho. Ao invés da ideia original de se juntar várias sondagens todas com o mesmo alcance regional (neste caso a nível de Portugal Continental) possibilitando desta forma o aumento da amostra quer a nível continental, quer a nível distrital, utilizaram-se sondagens realizadas num espaço temporal semelhante mas com um alcance geográfico mais reduzido (a nível distrital), conseguindo desta forma um aumento do tamanho de amostra em alguns dos distritos. A ideia era a de aproveitar recursos que poderiam já existir de forma natural num instituto de sondagens. Sondagens realizadas com propósitos diferentes mas com questões similares, realizadas segundo os mesmos processos e num mesmo espaço temporal poderiam ser aproveitadas para a obtenção de melhores estimativas a um nível micro.

No ano de 2005, a nível do continente, foi realizada uma sondagem cujo objectivo principal era a averiguação da intenção de voto dos eleitores para as eleições legislativas, sondagem descrita na secção 3.1. Existem ainda dados relativos a sondagens efectuadas nos distritos do Porto e Aveiro, também com objectivo principal de conhecer a intenção de voto nas eleições legislativas. Assim sendo pretende-se, recorrendo ao método de agregação de sondagens, utilizar a informação relativa a todas as sondagens (através da sua agregação) com vista a melhorar as previsões, quer a nível de distrito quer a nível de concelho. A Tabela 11 traduz a metodologia aplicada.

Tabela 11 – Descrição espaço-temporal das sondagens utilizadas.

Distrito	Jan - 05	Fev - 05	Fev - 05
	Objectivo Principal: Intenção de voto nas Legislativas	Objectivo Principal: Intenção de voto nas Legislativas	Objectivo Principal: Intenção de voto nas Legislativas
Aveiro	x	x	
Beja		x	
Braga		x	
Bragança		x	
Castelo Branco		x	
Coimbra		x	
Évora		x	
Faro		x	
Guarda		x	
Leiria		x	
Lisboa		x	
Portalegre		x	
Porto		x	x
Santarém		x	
Setúbal		x	
Viana do Castelo		x	
Vila Real		x	
Viseu		x	

No final de Janeiro de 2005 foi realizada uma sondagem a nível do distrito de Aveiro com o objectivo principal de se conhecer a intenção de voto para as eleições legislativas. Em Fevereiro foram realizadas duas sondagens com o mesmo objectivo, uma a nível do continente, outra a nível do distrito do Porto. A finalidade é melhorar as previsões a nível de distrito (e a nível de continente) através da agregação, dessas sondagens.

Os resultados das sondagens auxiliares (sondagens a nível do distrito do Porto e de Aveiro), após a eliminação dos elementos que não revelaram a sua intenção de voto ou optaram pelas opções “Indeciso” e “Não vai votar”, recodificados à semelhança do descrito na secção 3.1 para a sondagem a nível do continente, estão apresentados nas Tabelas 12 e 13, em que se apresentam as frequências absolutas (Freq. Abs.) e relativas (Freq. Rel.).

Tabela 12 – Resultado da sondagem a nível do distrito do Porto.

Partido	Freq Abs.	Freq. Rel.
PS	336	0,4941
Outro	344	0,5059
Total	680	1

Tabela 13 – Resultado da sondagem a nível do distrito de Aveiro.

Partido	Freq Abs.	Freq. Rel.
PS	247	0,4442
Outro	309	0,5558
Total	556	1

Agregando as três sondagens referidas anteriormente obtêm-se as estimativas e os coeficientes de variação apresentados na Tabela 14.

Tabela 14 – Resultados da agregação de sondagens

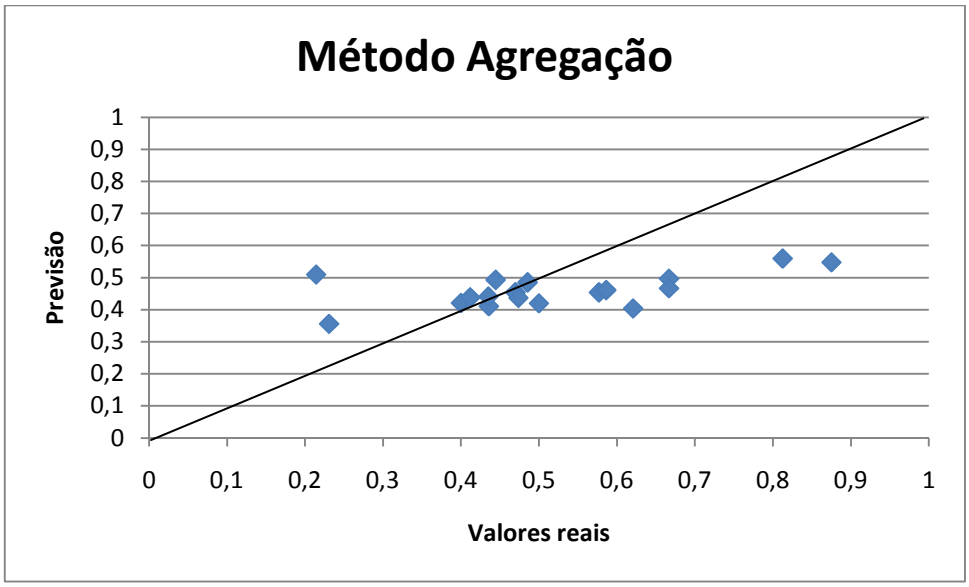
Distrito	Freq. Rel. PS	Freq. Rel. Outro	Peso	CV PS	CV Outro
Aveiro	**0,4357**	**0,5643**	**7,13%**	**1,14**	**0,88**
Beja	0,2143	0,7857	1,68%	1,92	0,52
Braga	0,4697	0,5303	8,07%	1,06	0,94
Bragança	0,4000	0,6000	1,55%	1,22	0,82
Castelo Branco	0,8125	0,1875	2,18%	0,48	2,08
Coimbra	0,5769	0,4231	4,56%	0,86	1,17
Évora	0,6667	0,3333	1,80%	0,71	1,42
Faro	0,4444	0,5556	4,06%	1,12	0,89
Guarda	0,6667	0,3333	1,87%	0,71	1,42
Leiria	0,2308	0,7692	4,65%	1,82	0,55
Lisboa	0,4351	0,5649	21,90%	1,14	0,88
Portalegre	0,8750	0,1250	1,32%	0,38	2,65
Porto	**0,4853**	**0,5147**	**17,65%**	**1,03**	**0,97**
Santarém	0,5862	0,4138	4,69%	0,84	1,19
Setúbal	0,4737	0,5263	8,05%	1,05	0,95
Viana do Castelo	0,5000	0,5000	2,56%	1,00	1,00
Vila Real	0,4118	0,5882	2,28%	1,19	0,84
Viseu	0,6207	0,3793	3,99%	0,78	1,28
Previsão Continente	0,4811	0,5190			

A Tabela 15 e o Gráfico 1 mostram a comparação entre as previsões obtidas através da aplicação do método de agregação de sondagens com os resultados reais das eleições de 2005.

Tabela 15 – Comparação método desagregação – resultados reais

Distrito	Freq. Rel. PS	Reais
Aveiro	**0,4357**	**0,4109**
Beja	0,2143	0,5101
Braga	0,4697	0,4542
Bragança	0,4000	0,4207
Castelo Branco	0,8125	0,5598
Coimbra	0,5769	0,4541
Évora	0,6667	0,4968
Faro	0,4444	0,4933
Guarda	0,6667	0,4670
Leiria	0,2308	0,3558
Lisboa	0,4351	0,4411
Portalegre	0,8750	0,5480
Porto	**0,4853**	**0,4853**
Santarém	0,5862	0,4614
Setúbal	0,4737	0,4371
Viana do Castelo	0,5000	0,4197
Vila Real	0,4118	0,4384
Viseu	0,6207	0,4040
Previsão Continente	0,4811	0,4514
Erro Abs Médio	0,1163	

Gráfico 1 – Comparação método desagregação – resultados reais



Como é visível pelas Tabelas 14 e 15, os resultados obtidos através da desagregação das sondagens melhoraram as previsões a nível de distrito (quer as considerando isoladamente a sondagem a nível continental, quer as considerando isoladamente as sondagens a nível distrital), mas pioraram as previsões a nível continental. As diferenças em relação aos resultados originais não são muito notórias devido à pouca quantidade de sondagens existentes para proceder à agregação, uma vez que só se possua sondagens auxiliares para os distritos do Porto e Aveiro (resultados a negrito nas tabelas) só foi possível aplicar este método nestes casos. Note-se contudo que o resultado obtido nestes distritos é consideravelmente melhor do que a estimativa fornecida apenas pelas sondagens, salientando-se o excelente resultado obtido no distrito do Porto.

Através do Gráfico 1 é possível verificar a diferença entre as previsões apresentadas pelo método de Agregação (e da sondagem, uma vez que excepto para o caso de Aveiro e Porto estas estimativas são as mesmas) e os valores reais observados nas eleições de 2005.

A não existência/indisponibilidade de dados, impossibilita uma maior exploração desta técnica, que, apesar de todos os perigos existentes na sua aplicação, nos deixa com a ideia de que poderá ser uma alternativa válida em casos em que existam sondagens disponíveis, não necessariamente realizadas com o mesmo fim, mas com a mesma questão, e com a condicionante de não distem muito no tempo, nem na metodologia de recolha de informação.

3.3 - Aplicação do Método de Regressão Multinível

Nesta secção apresentam-se os resultados obtidos através da aplicação do método de regressão multinível. Com o objectivo de encontrar o melhor modelo para o ajuste dos dados foram definidos quatro modelos de regressão multinível fazendo uso das variáveis demográficas e da interacção entre elas. Começou-se por definir um modelo mais complexo, utilizando todas as variáveis disponíveis e a interacção entre elas. Posteriormente, e como nem sempre o modelo mais complexo é o melhor, passou-se para a definição de modelos mais simples não recorrendo à utilização de interacções ou mesmo não recorrendo à utilização de todas as variáveis. As dificuldades encontradas na convergência dos modelos fizeram com que se optasse por definir um último modelo (Modelo 4) em que se faz recurso da utilização de parâmetros redundantes para acelerar a convergência.

Modelo 1

Recorde-se de que a intenção de voto, y_i, encontra-se codificada como 1 para PS e 0 para outros. Definiu-se um primeiro modelo, considerando as respostas independentes e utilizando todas as variáveis demográficas e as suas interacções. Como refere Gelman et al (2003), não há vantagens em fazer uma modelação multinível das variáveis com menos de 3 níveis, como o sexo ou a profissão, quando as distribuições a priori são não informativas. Desta forma o modelo definido tomando por base o modelo (2.33), considerando a idade, a escolaridade, as interacções existentes entre estas variáveis e o distrito como coordenadas na origem variáveis é:

$$\begin{aligned} \mathrm{P_r}(\mathrm{y_i}=1) = \mathrm{logit}^{-1}\big(&\beta^0 + \beta^{\mathrm{sexo}}.\,\mathrm{sexo_i} + \beta^{\mathrm{prof}}.\,\mathrm{prof_i} + \beta^{\mathrm{sexo.prof}}.\,\mathrm{sexo_i}.\,\mathrm{prof_i} \\ &+\propto^{\mathrm{idade}}_{\mathrm{k[i]}} +\propto^{\mathrm{esc}}_{\mathrm{l[i]}} +\propto^{\mathrm{idade.esc}}_{\mathrm{k[i],l[i]}} +\propto^{\mathrm{dis}}_{\mathrm{j[i]}}\big), \text{para i} = 1,\dots,719 \end{aligned} \tag{3.1}$$

$$\propto^{\mathrm{dis}}_{\mathrm{j[i]}} \sim \mathrm{N}\big(\beta^{\mathrm{v.prev}}, \sigma^2_{\mathrm{distrito}}\big) \tag{3.2}$$

$$\alpha^{\mathrm{idade}}_{\mathrm{k}} \sim \mathrm{N}\big(0, \sigma^2_{\mathrm{idade}}\big), \text{para k} = 1,..,6 \tag{3.3}$$

$$\alpha^{\mathrm{esc}}_{\mathrm{k}} \sim \mathrm{N}(0, \sigma^2_{\mathrm{esc}}), \text{para l} = 1,..,8 \tag{3.4}$$

$$\alpha^{\mathrm{idade.esc}}_{\mathrm{k,l}} \sim \mathrm{N}\big(0, \sigma^2_{\mathrm{idade.esc}}\big), \text{para k} = 1{,}6, \mathrm{l} = 1,..,8 \tag{3.5}$$

Em que β^0 é o termo constante para a coordenada na origem, β^{sexo} e β^{prof} denotam os coeficientes para as variáveis sexo e profissão, respectivamente; $sexo_i$ representa o sexo do indivíduo *i*, $prof_i$ é o indicador para o exercício de profissão do individuo *i*, $\beta^{sexo.prof}$ é o coeficiente para a interacção entre as variáveis sexo e profissão, $\propto_{k[i]}^{idade}$, $\propto_{l[i]}^{esc}$, $\propto_{j[i]}^{dis}$ e $\propto_{k[i],l[i]}^{idade.esc}$ são os coeficiente para a coordenada na origem das variáveis idade, escolaridade, distrito e da interacção entre as variáveis idade e escolaridade, respectivamente; e em que v.prev é a média dos resultados obtidos pelo PS nas duas anteriores eleições. As variâncias das variáveis distrito, idade, escolaridade e da interacção entre idade e escolaridade estão denotadas, respectivamente, por $\sigma^2_{distrito}$, σ^2_{idade}, σ^2_{esc} e $\sigma^2_{idade.esc}$.

A definição de um parâmetro para a média da distribuição normal de $\propto_{j[i]}^{dis}$, dependente da média dos resultados eleitorais em anos anteriores, $\beta^{v.prev}$, é utilizada por Gelman et al (2007) e Park et al (2004). Apesar de uma definição deste tipo poder ter desvantagens, uma vez que as intenções de voto dos eleitores variam no espaço temporal, neste caso de aplicação, não é de esperar uma mudança extremamente brusca na opinião do eleitorado e, caso esta existisse, essa tendência seria notada pela sondagem realizada, pelo que também aqui será utilizada, salientando porém as suas desvantagens.

Seguindo a metodologia utilizada por Gelman et al (2007), começou-se por implementar este modelo em R, utilizando a função lmer(), que permite o ajuste de modelos mistos, lineares, generalizados ou não lineares a um conjunto de dados.

Figura 4 – Output R do modelo 1 com recurso à função lmer()

```
> display (M1)
glmer(formula = voto ~ sexo + prof + sexo:prof + v.prev + (1 |
    idade) + (1 | esc) + (1 | idade.esc) + (1 | dis), family = binomial(link = "logit"))
            coef.est coef.se
(Intercept) -3.87     1.42
sexo         0.47     0.48
prof         0.46     0.51
v.prev       7.33     2.87
sexo:prof   -0.28     0.31

Error terms:
 Groups    Name        Std.Dev.
 idade.esc (Intercept) 0.29
 dis       (Intercept) 0.31
 esc       (Intercept) 0.00
 idade     (Intercept) 0.00
 Residual              1.00
---
number of obs: 719, groups: idade.esc, 39; dis, 18; esc, 8; idade, 6
AIC = 997.5, DIC = 979.5
deviance = 979.5
```

Na Figura 5 são visíveis as estimativas para a média da coordenada na origem e os coeficientes (e os seus desvios padrão) para as variáveis sexo, profissão, bem como a interacção entre as duas variáveis. Encontra-se também a estimativa para $\sigma_{distrito}$, σ_{idade}, σ_{esc} e $\sigma_{idade.esc}$. O objectivo principal desta análise não é a estimação dos parâmetros, mas sim a estimação da percentagem média de votos no PS dentro de cada estrato e, consequentemente, a percentagem média de votos no PS a nível da população.

O valor da coordenada na origem não é facilmente interpretável, uma vez que corresponde ao caso em que sexo, profissão e v.prev é igual a zero.

Gelman et al (2007), apresentam como regra de conveniência para a interpretação dos parâmetros a divisão da estimativa do parâmetro por quatro. O coeficiente para sexo é 0,47, o que dividido por 4, fornece uma estimativa grosseira, depois de controladas as restantes variáveis, de que as mulheres sem profissão estão, ligeiramente, mais determinadas (11,75%) a votar PS do que os homens sem profissão. Contudo, como o erro padrão deste coeficiente é tão elevado como a estimativa em si, o tamanho da amostra não é suficientemente grande para se ter a certeza desta interpretação, o mesmo se passando para o caso da interacção entre sexo e profissão.

O coeficiente para v.prev é 7,33, que dividido por 4 dá 1,83, o que sugere que uma diferença de 1% na média dos votos no PS nas últimas eleições é mapeada numa diferença preditiva de 1,83% de apoio no PS.

Os erros a nível de distrito têm um desvio padrão estimado de 0,31, o que dividido por 4 nos diz que diferem aproximadamente $\pm 7{,}8\%$ na escala de probabilidade (acima e além das diferenças explicadas pelos factores demográficos). As diferenças entre os grupos idade-escolaridade são aproximadamente $\pm 7{,}3\%$ na escala de probabilidades. Pouca ou nenhuma variabilidade foi encontrada entre os grupos de idade e os grupos de escolaridade, depois de controladas as restantes variáveis.

De seguida, procedeu-se à implementação deste modelo em WinBUGS, de modo a se obter inferências e predições mais precisas, uma vez que, com tantos factores, a inferência aproximada pela função lmer() (que não tem em conta a incerteza da variância dos parâmetros) não é tão credível. O código utilizado a implementação dos modelos em WinBUGS foi adaptado do código utilizado por Gelman (2007).

Após correr o programa 120000 vezes, de modo a que o parâmetro $\hat{R}$ seja menor que 1.1, facto que indica a convergência do modelo (Gelman et al (2007)), obtiveram-se os resultados para as estimativas dos parâmetros que se encontram na primeira ilustração do Anexo II. Nesta ilustração encontram-se os intervalos obtidos para os parâmetros através das diferentes simulações.

Nas figuras nos Anexo II estão apresentados os outputs obtidos através das simulações para os diferentes modelos, onde são visíveis intervalos para as diferentes simulações dos parâmetros. Da análise dos intervalos para o Modelo I retira-se que, uma vez que todos os intervalos possuem no seu interior o 0, nenhuma das variáveis se destaca pela sua capacidade preditiva da intenção de voto.

De seguida, para obter a estimativa da proporção de elementos que vota PS a nível continental, recorreu-se a uma pós-estratificacação dos resultados obtidos através das simulações, utilizando, para tal, os dados populacionais dos Censos 2001 obtidos

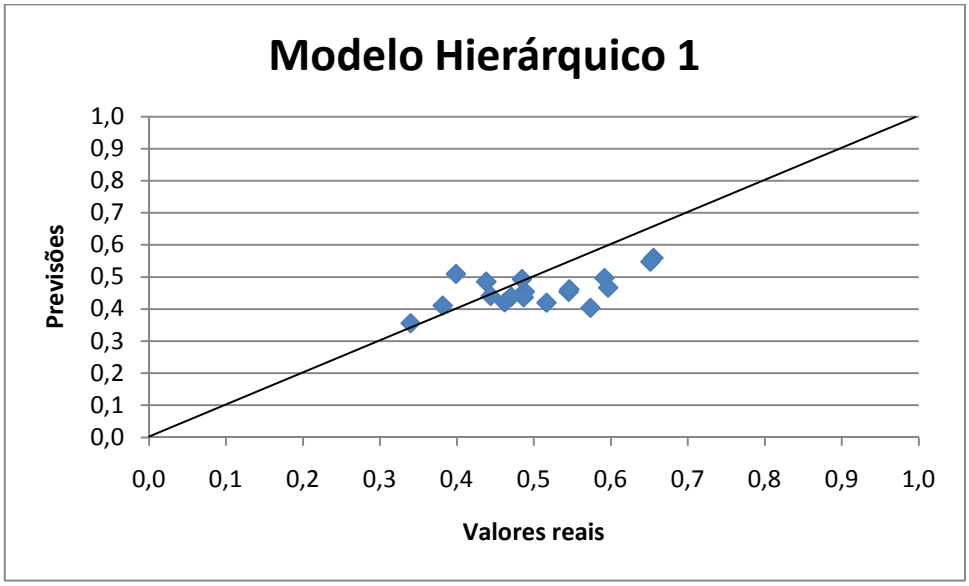
através do INE – Instituto Nacional de Estatística. De modo a estratificar a população em todas as variáveis escolhidas, juntamente com o distrito, necessita-se da distribuição conjunta da população das variáveis demográficas dentro de cada distrito. Como aproximação dessa distribuição utilizaram-se os dados dos Censos 2001. A fonte considerada possui a distribuição conjunta das variáveis demográficas dentro de cada distrito ponderadas para representar o total da população.

Após esta pós-estratificação, podemos sumariar as simulações realizadas para obter um ponto preditor (média das estimativas) e intervalos de incerteza, obtidos através da distribuição das simulações, para a proporção de adultos que votará PS, em cada distrito.

Tabela 16 – Comparação modelo hierárquico 1 – resultados reais.

Concelho	Quantil 0,25	Quantil 0,50	Quantil 0,75	Média	CV	Real
Aveiro	0,3225	0,3840	0,4360	0,3817	0,2191	0,4109
Beja	0,3283	0,4003	0,4744	0,3989	0,2738	0,5101
Braga	0,4349	0,4897	0,5416	0,4882	0,1535	0,4542
Bragança	0,3858	0,4607	0,5357	0,4622	0,2372	0,4207
Castelo Branco	0,5813	0,6587	0,7386	0,6555	0,1727	0,5598
Coimbra	0,4862	0,5477	0,6105	0,5453	0,1713	0,4541
Évora	0,5159	0,5866	0,6707	0,5919	0,1904	0,4968
Faro	0,4233	0,4827	0,5480	0,4846	0,1901	0,4933
Guarda	0,5188	0,5972	0,6724	0,5966	0,1907	0,4670
Leiria	0,2789	0,3357	0,4002	0,3400	0,2535	0,3558
Lisboa	0,4025	0,4438	0,4863	0,4439	0,1464	0,4411
Portalegre	0,5561	0,6542	0,7397	0,6514	0,1927	0,5480
Porto	0,3977	0,4395	0,4804	0,4382	0,1424	0,4853
Santarém	0,4853	0,5440	0,6124	0,5463	0,1734	0,4614
Setúbal	0,4349	0,4881	0,5403	0,4869	0,1583	0,4371
Viana do Castelo	0,4527	0,5188	0,5848	0,5166	0,1889	0,4197
Vila Real	0,4012	0,4713	0,5379	0,4705	0,2119	0,4384
Viseu	0,5193	0,5747	0,6356	0,5737	0,1544	0,4040
	Previsão Continente			0,4718		0,4514
	Erro Absoluto Médio			0,0688		
	DIC			990		

Gráfico 2 – Comparação modelo hierárquico 1 – resultados reais.



Pela análise da Tabela 16, que contém os intervalos de incerteza para as previsões, é possível verificar que o modelo 1 obtém um bom resultado a nível da previsão continental, bem como se assinalam algumas previsões a nível distrital muito satisfatórias. Salientam-se as previsões para Faro, Leiria e Lisboa, sendo ainda que 8 previsões para os distritos se encontram entre os quantis 0,25 e 0,75 das simulações (sombreadas a amarelo na Tabela 16). Note-se também no baixo erro absoluto médio (0,069) obtido através da modelação dos dados por este modelo. Este modelo possui um parâmetro DIC igual a 990.

Foram definidos mais três modelos, sendo que, para todos, a metodologia utilizada foi em tudo semelhante ao caso descrito até agora.

Modelo 2

Este modelo é semelhante ao Modelo 1, mas desta vez sem considerar as interacções entre as variáveis.

$$P_r(y_i = 1) = \text{logit}^{-1}\left(\beta^0 + \beta^{sexo}.\,sexo_i + \beta^{prof}.\,prof_i + \propto_{k[i]}^{idade} + \propto_{l[i]}^{esc} + \propto_{j[i]}^{dis}\right), \text{para } i = 1, \dots, 719 \quad (3.6)$$

$$\propto_{j[i]}^{dis} \sim N\left(\beta^{v.prev}, \sigma^2_{distrito}\right) \quad (3,7)$$

$$\alpha_k^{idade} \sim N\left(0, \sigma^2_{idade}\right), \text{para } k = 1, .., 6 \quad (3.8)$$

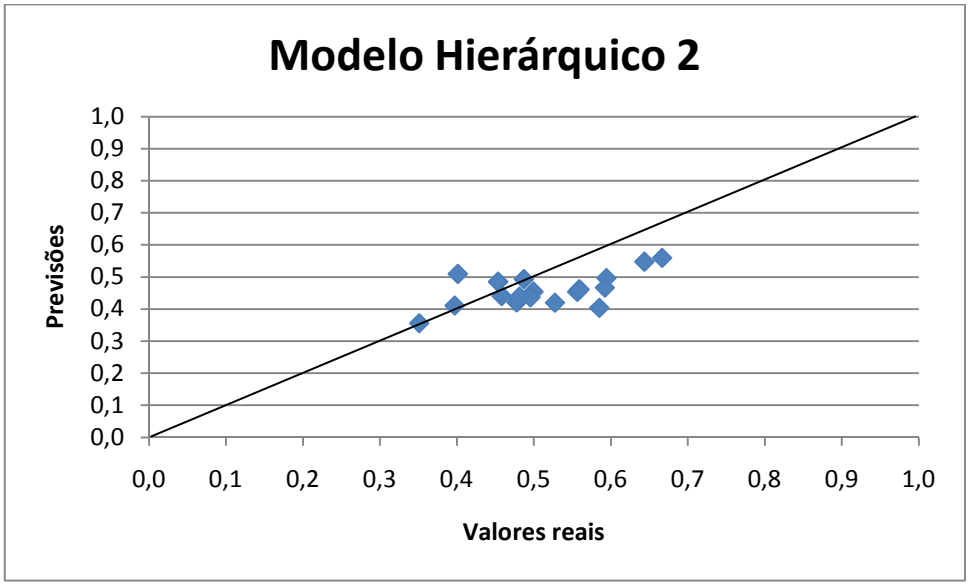
$$\alpha_k^{esc} \sim N(0, \sigma^2_{esc}), \text{para } l = 1, .., 8 \quad (3.9)$$

Saliente-se que os parâmetros são em tudo análogos à definição do Modelo 1. A estratégia aplicada para o modelo anterior foi repetida, sendo que este modelo obteve convergência, $\widehat{R} < 1.1$, com 180000 iterações.

Tabela 17 – Comparação modelo hierárquico 2 – resultados reais

Concelho	Quantil 0,25	Quantil 0,50	Quantil 0,75	Média	CV	Real
Aveiro	0,3462	0,3981	0,4466	0,3969	0,1769	0,4109
Beja	0,3257	0,4088	0,4802	0,4010	0,2674	0,5101
Braga	0,4623	0,5022	0,5350	0,4993	0,1123	0,4542
Bragança	0,4140	0,4802	0,5451	0,4775	0,2057	0,4207
Castelo Branco	0,5915	0,6738	0,7381	0,6663	0,1462	0,5598
Coimbra	0,5007	0,5563	0,6071	0,5561	0,1380	0,4541
Évora	0,5236	0,5947	0,6614	0,5941	0,1714	0,4968
Faro	0,4410	0,4867	0,5308	0,4869	0,1502	0,4933
Guarda	0,5210	0,5907	0,6613	0,5921	0,1709	0,4670
Leiria	0,2978	0,3464	0,4024	0,3509	0,2219	0,3558
Lisboa	0,4324	0,4589	0,4840	0,4580	0,0860	0,4411
Portalegre	0,5630	0,6396	0,7203	0,6434	0,1740	0,5480
Porto	0,4278	0,4534	0,4816	0,4533	0,0868	0,4853
Santarém	0,5065	0,5581	0,6068	0,5590	0,1337	0,4614
Setúbal	0,4557	0,4952	0,5360	0,4954	0,1184	0,4371
Viana do Castelo	0,4735	0,5235	0,5769	0,5272	0,1505	0,4197
Vila Real	0,4218	0,4824	0,5408	0,4809	0,1852	0,4384
Viseu	0,5276	0,5837	0,6371	0,5848	0,1328	0,4040
	Previsão Continente			0,4832		0,4514
	Erro Absoluto Médio			0,0721		
	DIC			989,1		

Gráfico 3 – Comparação modelo hierárquico 2 – resultados reais.



Pela análise dos resultados obtidos, da Tabela 17 e do Gráfico 3, pode-se concluir que este modelo obtém um resultado pior do que o modelo 1 a nível da previsão nacional, facto que se traduz no seu erro absoluto médio superior comparativamente ao caso modelo 1. Este modelo obtém contudo algumas boas previsões a nível de distrito, salientando-se os casos de Aveiro, Leiria e Lisboa. Note-se também que apenas 6 previsões a nível de distrito se encontram entre os quantis 0,25 e 0,75 das simulações. (sombreadas a amarelo na Tabela 17). Note-se também que o valor para o parâmetro DIC deste modelo (989,1) é muito semelhante ao obtido pelo Modelo 1 (990).

Modelo 3

Este modelo é mais simples do que os anteriores, considerando apenas as variáveis demográficas sexo e idade.

$$\mathrm{P_r}(\mathrm{y_i}=1)=\mathrm{logit}^{-1}\left(\beta^{0}+\beta^{\mathrm{sexo}}.\mathrm{sexo_i}+\propto_{\mathrm{k[i]}}^{\mathrm{idade}}+\propto_{\mathrm{j[i]}}^{\mathrm{dis}}\right),\text{para i}=1,\ldots,719 \quad (3.10)$$

$$\alpha_{\mathrm{k}}^{\mathrm{idade}}\sim\mathrm{N}\left(0,\sigma_{\mathrm{idade}}^{2}\right),\text{para k}=1,..,6 \quad (3.11)$$

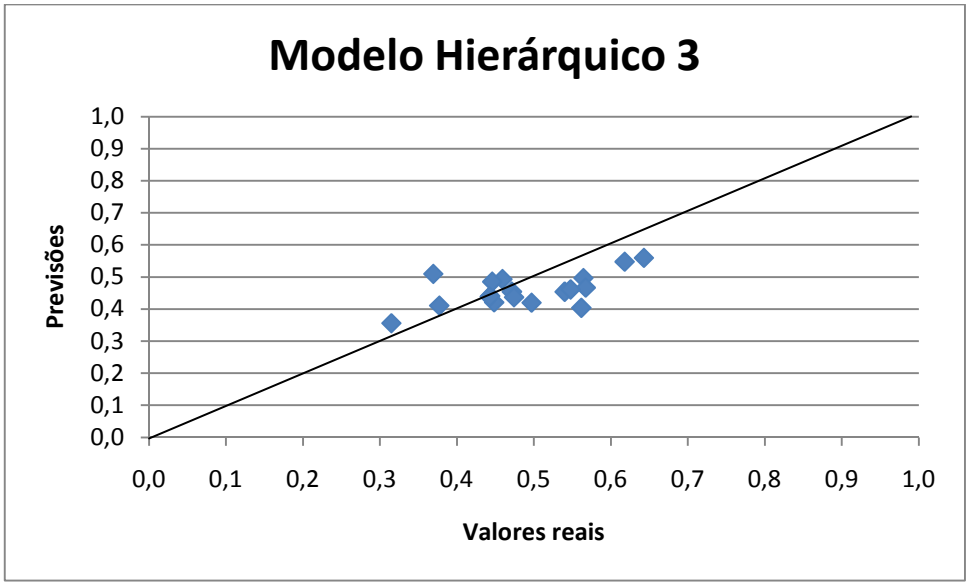
$$\propto_{\mathrm{j[i]}}^{\mathrm{dis}}\sim\mathrm{N}\left(\beta^{\mathrm{v.prev}},\sigma_{\mathrm{distrito}}^{2}\right) \quad (3.12)$$

Em que os parâmetros são análogos às definições anteriores. Para este modelo, a convergência, $\widehat{\mathrm{R}}<1.1$, foi atingida com 30000 iterações, o que é valor muito inferior ao número de interacções necessárias para a obtenção de convergência com os restantes modelos. Tal facto é devido à maior simplicidade do modelo.

Tabela 18 – Comparação modelo hierárquico 3 – resultados reais.

Concelho	Quantil 0,25	Quantil 0,50	Quantil 0,75	Média	CV	Real
Aveiro	0,3292	0,3775	0,4241	0,3770	0,1833	0,4109
Beja	0,3004	0,3719	0,4367	0,3693	0,2581	0,5101
Braga	0,4361	0,4710	0,5041	0,4714	0,1090	0,4542
Bragança	0,3810	0,4456	0,5165	0,4484	0,2171	0,4207
Castelo Branco	0,5678	0,6415	0,7196	0,6430	0,1611	0,5598
Coimbra	0,4889	0,5409	0,5924	0,5401	0,1436	0,4541
Évora	0,4998	0,5598	0,6234	0,5644	0,1721	0,4968
Faro	0,4082	0,4578	0,5093	0,4592	0,1619	0,4933
Guarda	0,4986	0,5660	0,6377	0,5674	0,1760	0,4670
Leiria	0,2626	0,3106	0,3657	0,3148	0,2220	0,3558
Lisboa	0,4187	0,4427	0,4679	0,4433	0,0853	0,4411
Portalegre	0,5367	0,6122	0,6962	0,6180	0,1781	0,5480
Porto	0,4207	0,4449	0,4736	0,4458	0,0857	0,4853
Santarém	0,4945	0,5475	0,5971	0,5477	0,1351	0,4614
Setúbal	0,4332	0,4741	0,5125	0,4742	0,1226	0,4371
Viana do Castelo	0,4437	0,4949	0,5503	0,4971	0,1560	0,4197
Vila Real	0,3890	0,4426	0,4979	0,4424	0,1875	0,4384
Viseu	0,5042	0,5581	0,6197	0,5617	0,1445	0,4040
	Previsão Continente			0,4638		0,4514
	Erro Absoluto Médio			0,0614		
	DIC			987,6		

Gráfico 4 – Comparação previsões hierárquico 3 – resultados reais.



Através da análise dos resultados deste modelo, da Tabela 18 e do Gráfico 4, verifica-se a existência de 10 previsões a nível de distrito que se encontram entre os quantis 0,25 e 0,75 das simulações (sombreadas a amarelo na Tabela 18). É também de salientar o seu baixo erro absoluto médio (0,061), bem como as previsões para Lisboa e Vila Real. Apesar de ser o modelo mais simples, este modelo é, até agora, o que apresenta os melhores resultados, possuindo também o valor para o parâmetro DIC mais baixo (987,6)

Modelo 4

Este modelo é semelhante ao Modelo 1, mas utilizando parâmetros redundantes na implementação do código computacional, de modo a tornar mais rápida a simulação, isto é possível acrescentando ao modelo coeficientes redundantes que são colineares com preditores existentes no modelo, seguindo a metodologia de Gelman et al (2007).

$$\mathrm{P_r}(y_i = 1) = \mathrm{logit}^{-1}\left(\beta^0 + \beta^{sexo}.\,sexo_i + \beta^{prof}.\,prof_i + \beta^{sexo.prof}.\,sexo_i.\,prof_i\right) \quad (3.13)$$
$$+\propto^{idade}_{k[i]} +\propto^{esc}_{l[i]} +\propto^{idade.esc}_{k[i],l[i]} +\propto^{dis}_{j[i]}\Big), \text{para } i = 1,\ldots,719$$

$$\propto^{dis}_{j[i]} \sim N\left(\beta^{v.prev}, \sigma^2_{distrito}\right) \quad (3.14)$$

$$\alpha^{idade}_{k} \sim N\left(\mu_{idade}, \sigma^2_{idade}\right), \text{para } k = 1,..,6 \quad (3.15)$$

$$\alpha^{esc}_{k} \sim N(\mu_{esc}, \sigma^2_{esc}), \text{para } l = 1,..,8 \quad (3.16)$$

$$\alpha^{idade.esc}_{k,l} \sim N\left(\mu_{idade.esc}, \sigma^2_{idade.esc}\right), \text{para } k = 1,6, l = 1,..,8 \quad (3.17)$$

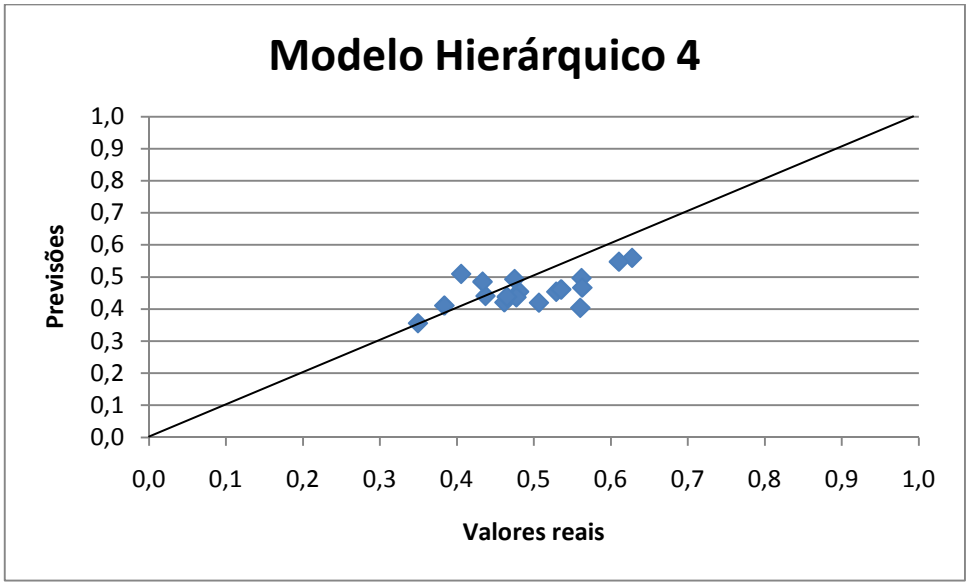
Em que, aos parâmetros μ_{idade}, μ_{esc} e $\mu_{idade.esc}$, são atribuídas distribuições *a priori* não informativas e os restantes são em tudo análogos às definições anteriores.

A convergência, $\widehat{R} < 1.1$, deste modelo foi obtida com 100000 iterações, o que, comparativamente ao modelo 1, representa uma diminuição de 20000 iterações.

Tabela 19 – Comparação modelo hierárquico 4 – resultados reais.

Concelho	Quantil 0,25	Quantil 0,50	Quantil 0,75	Média	CV	Real
Aveiro	0,3248	0,3840	0,4401	0,3836	0,2094	0,4109
Beja	0,3338	0,4100	0,4874	0,4054	0,2633	0,5101
Braga	0,4318	0,4812	0,5280	0,4807	0,1461	0,4542
Bragança	0,3907	0,4588	0,5355	0,4614	0,2328	0,4207
Castelo Branco	0,5464	0,6274	0,7071	0,6273	0,1774	0,5598
Coimbra	0,4644	0,5253	0,5895	0,5288	0,1697	0,4541
Évora	0,4850	0,5562	0,6406	0,5616	0,1921	0,4968
Faro	0,4127	0,4716	0,5343	0,4746	0,1910	0,4933
Guarda	0,4833	0,5510	0,6362	0,5627	0,1964	0,4670
Leiria	0,2905	0,3463	0,4075	0,3493	0,2428	0,3558
Lisboa	0,3954	0,4354	0,4771	0,4370	0,1403	0,4411
Portalegre	0,5240	0,6044	0,6928	0,6102	0,1981	0,5480
Porto	0,3892	0,4313	0,4696	0,4331	0,1407	0,4853
Santarém	0,4718	0,5323	0,5946	0,5352	0,1714	0,4614
Setúbal	0,4259	0,4753	0,5256	0,4771	0,1581	0,4371
Viana do Castelo	0,4414	0,5021	0,5685	0,5063	0,1886	0,4197
Vila Real	0,3959	0,4654	0,5276	0,4647	0,2098	0,4384
Viseu	0,4984	0,5593	0,6213	0,5601	0,1603	0,4040
	Previsão Continente			0,4637		0,4514
	Erro Absoluto Médio			0,0571		
	DIC			-3,14*10^12		

Gráfico 5 – Comparação modelo hierárquico 4 – resultados reais.



De todos os modelos, este é o que parece apresentar melhores resultados, este facto é visível no seu baixo erro absoluto médio (0,057) e na existência de 11 previsões a nível de distrito que se encontram entre os quantis 0,25 e 0,75 das simulações (sombreadas a amarelo na Tabela 19). Para além disso é o que possui o menor valor para o parâmetro DIC (-3,14*10^12, este parâmetro pode ser negativo uma vez que a função densidade de probabilidade, utilizada no cálculo do "desvio" pode ser maior do que 1, no caso de possuir um desvio padrão pequeno, segundo Spiegelhalter, D, (2006)). Pela análise do Gráfico 5 é possível verificar que as previsões fornecidas pelo modelo 4 não se afastam de forma expressiva dos valores reais observados.

Para análise das Tabelas 16, 17, 18 e 19 é possível verificar que todos os modelos apresentam coeficientes de variação baixos.

A Tabela 20 apresenta uma comparação das previsões obtidas através dos diferentes modelos, onde são visíveis as melhorias alcançadas através da aplicação dos modelos de regressão hierárquicos, relativamente às previsões fornecidas pela consideração isolada da sondagem. Salientam-se os casos dos distritos de Beja, Castelo Branco, Évora, Leiria e Portalegre, sendo que alguns destes distritos são mesmo os que possuem menor tamanho de amostra na sondagem, como é exemplo Portalegre.

A análise dos resultados obtidos salienta o Modelo 4 como a melhor opção, de entre os modelos definidos, para a melhoria das estimativas, quer a nível distrital, quer a nível de continente. Esta conclusão é fundamentada no menor erro obtido a nível da previsão continental, na presença do maior número de distritos dentro do intervalo de incerteza, encontrado através de simulações. Para além destas razões é, ainda, o modelo que apresenta o menor valor para o parâmetro DIC.

Tabela 20 – Comparação modelos hierárquicos – resultados reais.

	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Sondagem	Real
Aveiro	0,3817	0,3969	0,3770	0,3836	0,3256	0,4109
Beja	0,3989	0,4010	0,3693	0,4054	0,2143	0,5101
Braga	0,4882	0,4993	0,4714	0,4807	0,4697	0,4542
Bragança	0,4622	0,4775	0,4484	0,4614	0,4000	0,4207
Castelo Branco	0,6555	0,6663	0,6430	0,6273	0,8125	0,5598
Coimbra	0,5453	0,5561	0,5401	0,5288	0,5769	0,4541
Évora	0,5919	0,5941	0,5644	0,5616	0,6667	0,4968
Faro	0,4846	0,4869	0,4592	0,4746	0,4444	0,4933
Guarda	0,5966	0,5921	0,5674	0,5627	0,6667	0,4670
Leiria	0,3400	0,3509	0,3148	0,3493	0,2308	0,3558
Lisboa	0,4439	0,4580	0,4433	0,4370	0,4351	0,4411
Portalegre	0,6514	0,6434	0,6180	0,6102	0,8750	0,5480
Porto	0,4382	0,4533	0,4458	0,4331	0,4412	0,4853
Santarém	0,5463	0,5590	0,5477	0,5352	0,5862	0,4614
Setúbal	0,4869	0,4954	0,4742	0,4771	0,4737	0,4371
Viana do Castelo	0,5166	0,5272	0,4971	0,5063	0,5000	0,4197
Vila Real	0,4705	0,4809	0,4424	0,4647	0,4118	0,4384
Viseu	0,5737	0,5848	0,5617	0,5601	0,6207	0,4040
Previsão Continente	0,4718	0,4832	0,4638	0,4637	0,4617	0,4514
Erro absoluto médio	0,0688	0,0721	0,0614	0,0571	0,1221	
DIC	990	989,1	987,6	-3,14*10^12		

3.2 - Aplicação do EBLUP

Nesta secção apresentam-se os resultados obtidos através da aplicação do método EBLUP. Foram definidos dois modelos, todos tomando como base a sondagem eleitoral utilizada na aplicação dos métodos descritos anteriormente e, como informação auxiliar, um utiliza os resultados eleitorais nas eleições de 1999, outro utiliza os resultados eleitorais nas eleições de 2002. A ideia inicial era a de utilizar outras sondagens, realizadas como mesmo alcance geográfico, para servirem de informação auxiliar na formulação dos modelos, no entanto, tal não foi contudo possível devido à inexistência/inacessibilidade das mesmas. Optou-se então por utilizar os resultados eleitorais, numa perspectiva exploradora e com a precaução de se saber de antemão que a intenção de voto muda consoante o tempo.

Como se pode observar no Anexo III, o modelo utilizando como informação auxiliar os resultados eleitorais das eleições de 2002, considerando o modelo (2.16), pode ser definido da seguinte forma:

$$y_i = 0{,}3067534 + 0{,}1398341x_i \tag{3.18}$$

Em que y_i é a previsão dada pelo modelo da proporção de votos no PS no distrito i, e em que x_i é a previsão obtida pela sondagem para o mesmo distrito.

O modelo que utiliza como informação auxiliar os resultados eleitorais das eleições de 1999, considerando o modelo (2.16), pode ser definido como:

$$y_i = 0{,}368645 + 0{,}1479796x_i \tag{3.19}$$

Em que y_i é a previsão dada pelo modelo da proporção de votos no PS no distrito i, e em que x_i é a previsão obtida pela sondagem para o mesmo distrito.

Os resultados obtidos através da aplicação destes modelos estão representados nas Tabelas 21 e 22. O modelo obtido utilizando como informação auxiliar os resultados eleitorais de 1999 é o que obtém os melhores resultados, este facto deve-se à existência de tendências de voto similares nestas duas eleições, uma vez que a intenção de voto em

1999 era predominantemente PS (à semelhança de 2005) enquanto em 2002 era predominantemente PSD (ao contrário de 2005).

Os coeficientes de variação obtidos são muito similares dentro de cada modelo, embora sejam relativamente mais baixos no caso do modelo com informação auxiliar de 1999, o que demonstra haver menos dispersão neste modelo.

Note-se que o modelo que faz recurso de informação auxiliar relativa aos resultados eleitorais do ano de 2002, devido às diferenças nas intenções de voto entre as eleições de 2002 e de 2005, obtém resultados bastante imprecisos, quer a nível de distrito, quer a nível continental, dados comprovados pelo seu erro absoluto médio (0,0809), que, embora seja inferior ao da sondagem, não permite um aumento relevante na segurança das previsões.

Pela análise da aplicação e dos resultados obtidos através deste método fica a ideia de extrema sensibilidade face à informação auxiliar existente. Caso esta seja semelhante aos dados a modelar, o modelo terá um bom ajuste ao conjunto de dados e obterá bons resultados como preditor em pequenos domínios. Caso a informação auxiliar seja bastante diferente dos dados a modelar, aumentam muito as hipóteses de um mau ajuste ao conjunto de dados e obtenção de maus resultados como preditor.

A escolha da informação auxiliar para a aplicação deste método deve, portanto, ser feita de forma muito cautelosa, sendo que a escolha de informação semelhante ao conjunto de dados se traduzir em bons resultados enquanto a escolha de informação muito díspar dos dados em estudo se reflectirá provavelmente em maus resultados.

Através dos Gráficos 6 e 7 é possível verificar as diferenças entre os dois modelos definidos. As previsões do modelo com informação auxiliar de 2002 (Gráfico 6) encontram-se muito deslocadas dos valores reais, enquanto as previsões do modelo com informação auxiliar de 1999 (Gráfico 7) estão mais aderentes em relação aos reais valores observados.

Tabela 21 – Comparação modelo EBLUP 2002 – resultados reais.

Estimativas com base na Inf Aux 2002				
Distrito	**Freq. Rel PS**	**Reais**	**Diferença**	**CV**
Aveiro	0,3523	0,4109	-0,0586	1,4161
Beja	0,3367	0,5101	-0,1734	1,4816
Braga	0,3724	0,4542	-0,0818	1,3395
Bragança	0,3627	0,4207	-0,0580	1,3755
Castelo Branco	0,4204	0,5598	-0,1394	1,1868
Coimbra	0,3874	0,4541	-0,0667	1,2877
Évora	0,4000	0,4968	-0,0968	1,2473
Faro	0,3689	0,4933	-0,1244	1,3524
Guarda	0,4000	0,4670	-0,0670	1,2473
Leiria	0,3390	0,3558	-0,0168	1,4715
Lisboa	0,3676	0,4411	-0,0735	1,3572
Portalegre	0,4291	0,5480	-0,1189	1,1626
Porto	0,3684	0,4853	-0,1169	1,3540
Santarém	0,3887	0,4614	-0,0727	1,2834
Setúbal	0,3730	0,4371	-0,0641	1,3375
Viana do Castelo	0,3767	0,4197	-0,0430	1,3245
Vila Real	0,3643	0,4384	-0,0741	1,3693
Viseu	0,3935	0,4040	-0,0105	1,2677
Previsão Continente	0,3718	0,4514		
Erro Abs Médio	0,0809			

Gráfico 6 – Comparação modelo EBLUP 2002 – resultados reais.

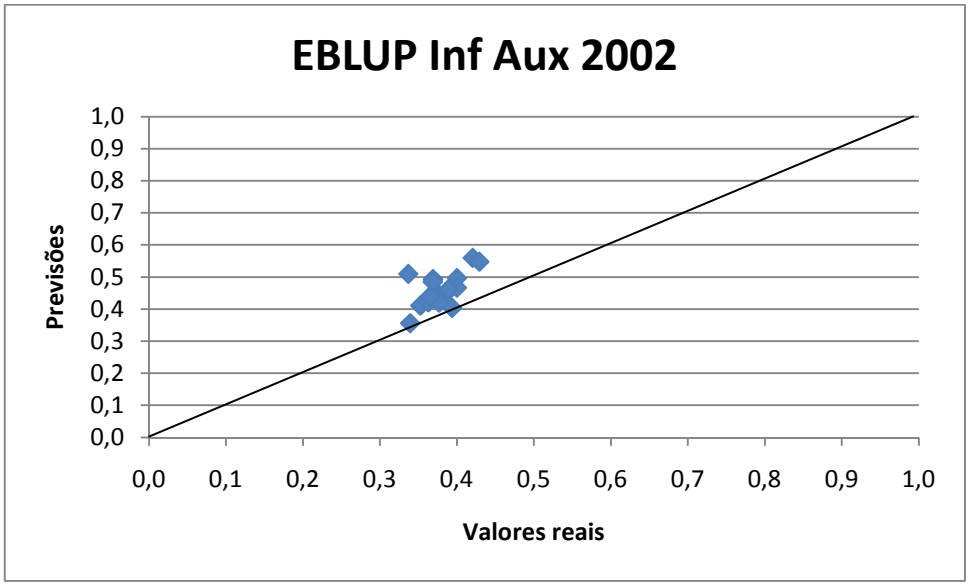
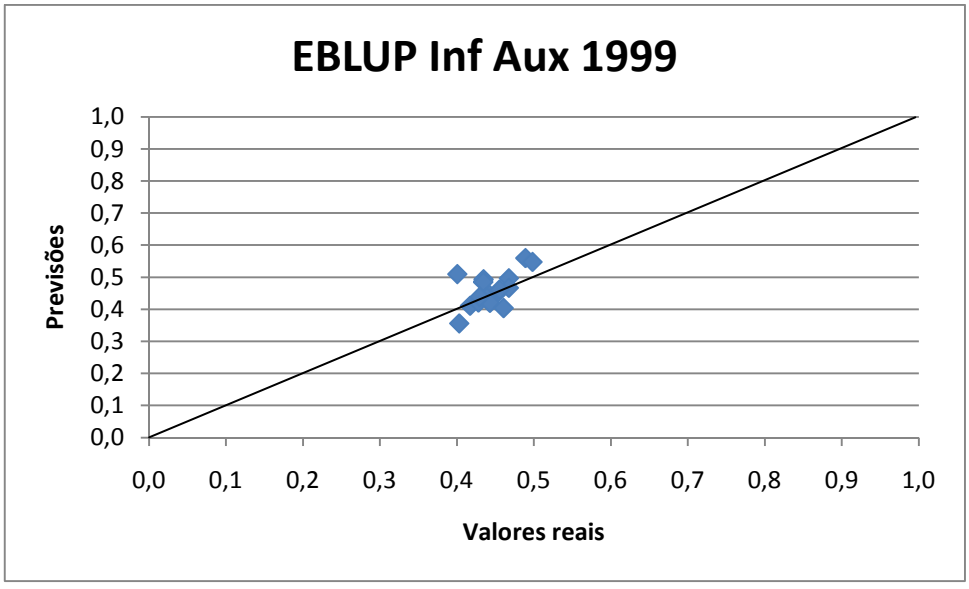


Tabela 22 – Comparação modelo EBLUP 1999 – resultados reais.

Estimativas com base na Inf Aux 1999				
Distrito	**Freq. Rel PS**	**Reais**	**Diferença**	**CV**
Aveiro	0,4169	0,4109	0,0060	1,1967
Beja	0,4004	0,5101	-0,1097	1,2460
Braga	0,4382	0,4542	-0,0160	1,1385
Bragança	0,4279	0,4207	0,0072	1,1659
Castelo Branco	0,4889	0,5598	-0,0709	1,0204
Coimbra	0,4541	0,4541	0,0000	1,0987
Évora	0,4673	0,4968	-0,0295	1,0675
Faro	0,4344	0,4933	-0,0589	1,1483
Guarda	0,4673	0,467	0,0003	1,0675
Leiria	0,4028	0,3558	0,0470	1,2384
Lisboa	0,4331	0,4411	-0,0080	1,1520
Portalegre	0,4982	0,548	-0,0498	1,0014
Porto	0,4340	0,4853	-0,0513	1,1496
Santarém	0,4554	0,4614	-0,0060	1,0954
Setúbal	0,4388	0,4371	0,0017	1,1370
Viana do Castelo	0,4427	0,4197	0,0230	1,1270
Vila Real	0,4296	0,4384	-0,0088	1,1612
Viseu	0,4605	0,404	0,0565	1,0833
Previsão Continente	0,4375	0,4514		
Erro Abs Médio	0,0306			

Gráfico 7 – Comparação modelo EBLUP 1999 – resultados reais.



Capítulo 4 - Conclusões

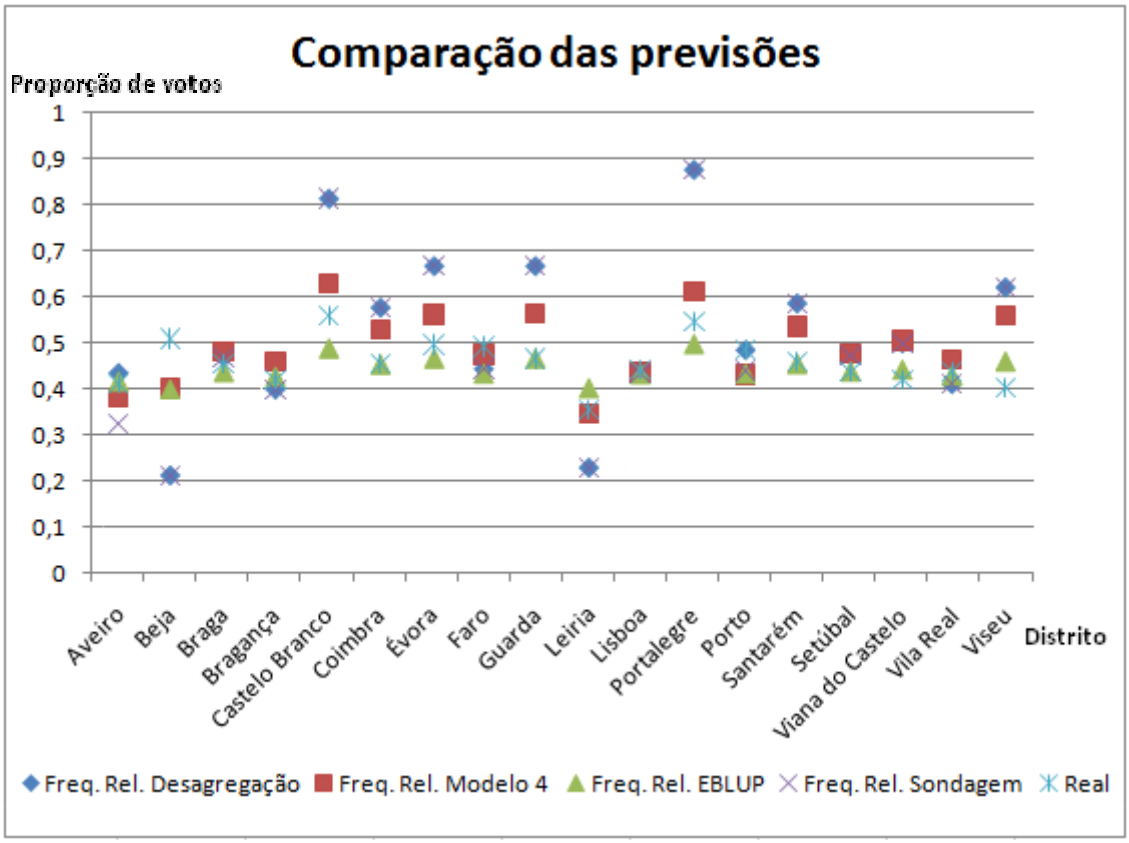
Neste capítulo faz-se a apresentação dos resultados obtidos e das conclusões, comparando os melhores de todos os métodos considerados.

Como é visível na Tabela 23, o método EBLUP com recurso à informação auxiliar dos resultados eleitorais de 1999 é o que apresenta melhores resultados a nível de distrito, uma vez que este método apresenta a melhor previsão em 11 dos 18 distritos (na Tabela 23 a amarelo), salientando-se o resultado de Coimbra. A nível do continente, o método que obtém melhores resultados é o Modelo Hierárquico 4 (na Tabela 23 a amarelo).

Tabela 23– Comparativo entre os melhores métodos.

	Freq. Rel. Desagregação	**Freq. Rel. Modelo 4**	**Freq. Rel. EBLUP 1999**	**Freq. Rel. Sondagem**	**Real**
Aveiro	**0,4357**	0,3836	0,4169	0,3256	0,4109
Beja	0,2143	0,4054	0,4004	0,2143	0,5101
Braga	0,4697	0,4807	0,4382	0,4697	0,4542
Bragança	0,4000	0,4614	0,4279	0,4000	0,4207
Castelo Branco	0,8125	0,6273	0,4889	0,8125	0,5598
Coimbra	0,5769	0,5288	0,4541	0,5769	0,4541
Évora	0,6667	0,5616	0,4673	0,6667	0,4968
Faro	0,4444	0,4746	0,4344	0,4444	0,4933
Guarda	0,6667	0,5627	0,4673	0,6667	0,4670
Leiria	0,2308	0,3493	0,4028	0,2308	0,3558
Lisboa	0,4351	0,4370	0,4331	0,4351	0,4411
Portalegre	0,8750	0,6102	0,4982	0,8750	0,5480
Porto	**0,4853**	0,4331	0,4340	0,4412	0,4853
Santarém	0,5862	0,5352	0,4554	0,5862	0,4614
Setúbal	0,4737	0,4771	0,4388	0,4737	0,4371
Viana do Castelo	0,5000	0,5063	0,4427	0,5000	0,4197
Vila Real	0,4118	0,4647	0,4296	0,4118	0,4384
Viseu	0,6207	0,5601	0,4605	0,6207	0,4040
Previsão Continente	0,4811	0,4637	0,4375	0,4655	0,4514
Erro absoluto médio	0,1163	0,0571	0,0306	0,1221	

Gráfico 8 – Comparativo entre as previsões dos melhores métodos



Analisando o Gráfico 8 é possível verificar as melhorias obtidas através da aplicação dos métodos nos distritos com menor tamanho de amostra na sondagem (por exemplo Castelo Branco e Portalegre).

O método de agregação de sondagens, pela forma como foi aplicado, só produz resultados práticos em dois distritos (a negrito na Tabela 23). Salienta-se contudo o resultado obtido no distrito do Porto, que fortalece a ideia de que este método pode ser uma alternativa válida quando aplicado com cautela relativamente aos métodos de amostragem das diferentes sondagens e espaços temporais de realização.

O método de regressão multinível apresenta-se como uma opção segura, um vez que obtém resultados muito satisfatórios utilizando como informação auxiliar pouco mais do que os dados demográficos da população. Este método consegue melhorar de forma relevante as previsões da sondagem a nível de distrito, facto que é sustentado pela redução do erro absoluto médio de 0,1221 para 0,571 e ainda por obter uma melhoria a nível da previsão para o continente.

Os excelentes resultados obtidos através do EBLUP devem ser encarados e interpretados com precaução, uma vez que tais são obtidos à custa de informação auxiliar que neste caso era semelhante àquela que se pretendia estimar. Poderemos nem sempre estar na presença de uma situação similar ou, com piores consequências, poderemos estar numa situação de falsa aparência de similaridade com a informação auxiliar, o que poderá se poderá traduzir em resultados verdadeiramente catastróficos.

Foi possível verificar que a aplicação das técnicas da agregação de sondagens e do EBLUP devem ser utilizadas com precaução no que se refere à utilização da informação auxiliar. Esta informação é de extrema importância para um bom ajuste destes modelos ao conjunto de dados a modelar, bem como para a obtenção de boas previsões. A escolha da informação adequada poderá resultar numa melhoria substancial nas estimativas, enquanto a escolha de informação desajustada poderá traduzir-se numa pioria de resultados em relação às previsões fornecidas por uma sondagem tomada de modo isolado. Como tal, esta escolha de informação deve ser feita, pelo estatístico, de forma conscienciosa, devidamente fundamentada e atenta a qualquer pormenor que possa incutir uma alteração na tendência de voto. Os bons resultados obtidos e a ausência de apreensões relativas à aplicação da metodologia e à escolha da informação auxiliar, sugerem que a técnica de regressão multinível pode e deve ser encarada como uma alternativa segura para a obtenção de estimativas mais precisas, quer a nível micro, quer a nível macro.

Através da aplicação destes métodos foi possível encontrar soluções válidas e eficientes para a melhoria das estimativas a nível micro, bem como a nível macro. Os resultados obtidos pela aplicação das metodologias evidenciam a capacidade de se melhorar as estimativas, a nível micro e a nível macro, obtidas através de uma sondagem, utilizando técnicas estatísticas, sendo de lamentar a impossibilidade de uma maior exploração do método de agregação e a exploração do EBLUP utilizando outro tipo de informação auxiliar. Estes métodos devem ser também encarados como alternativas válidas para a melhoria das estimativas noutras áreas para além das sondagens eleitorais, quando utilizados de forma adequada e de acordo com os seus pressupostos e metodologias de aplicação.

Capítulo 5 – Bibliografia

· Aaker, Kumar, Day, (2006). *Marketing Research*, 9/e, Wiley.

· ABS (2005), A Guide to Small Area Estimation - Executive Summary, Australian Bureau of Statistics,

· Anderson, Hair, Babin, Tathan, Black, (2006). *Multivariate Data Analysis, Sixth Edition*, Pearson Prentice Hall.

· Cochran, W.G., (1977). *Sampling Technique*, 3rd ed., Willey.

· Coelho, P., (1998), Estimadores combinados de pequenos domínios, in *Revista de Estatística*, 2o quad. de 1998, pp., INE

· Erikson, Robert S., Gerald C.Wright, and John P.McIver. (1993), *Statehouse Democracy: Public Opinion and Policy in the American States*. Cambridge: Cambridge University Press.

· Gelman, Andrew, and Jennifer Hill. (2007), *Data Analysis Using Regression and Multilevel Hierarchical Models*. Cambridge: Cambridge University Press.

· Gelman, Andrew, and Thomas C. Little. (1997), Poststratication into Many Categories Using Hierarchical Logistic Regression, in *Survey Methodology* 23(2): 127-35.

· Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. (2003). *Bayesian Data Analysis*, 2nd ed. London: Chapman and Hall.

· Gómez-Rubio, V., Bivand, R. S., Pebesma, E. (2008). *Applied Spatial Data Analysis With R*, Springer.

· Hill, A., Hill, M., (2005). *Investigação Por Questionário*, 2.ª Edição, Edições Sílabo.

· Kreft, I., Leeuw, J. (1998), *Introducing Multilevel Modeling*, Sage Publications.

· Kish, L., (1965). *Survey Sampling*, Willey.

· Lax, Phillips (2009), How Should We Estimate Public Opinion in The States?, in *American Journal of Political Science*, Vol. 53, No. 1, January 2009, Pp. 107-121

· Leeuw, J., Meijer, E. (Editores) (2008), *Handbook of Multilevel Analysis*, Springer

· Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**:325--337, http://www.mrc-bsu.cam.ac.uk/bugs/, acedido em 11 de Outubro de 2010.

· Maroco, J. (2007). *Análise Estatística com utilização do SPSS*, 3ª Edição, Edições Sílabo.

· Park, D., Gelman, A. And Bafumi, J. (2004), Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls, in *Political Analysis* (2004) 12:375–385.

· Pestana, M., Gageiro, J. (2005). *Análise de Dados para Ciências Sociais*, 4ª Edição. Edições Sílabo.

· R Development Core Team (2010). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, http://www.r-project.org/index.html, acedido em 11 de Outubro de 2010.

· Rao, C.R. (2003). *Small Area Estimation*. New Jersey: John Wiley and Sons.

· Reis, E., Ferrão, F., Vicente, P. (2001). *Sondagens. A amostragem como factor decisivo de qualidade*, 2.ª Edição, Edições Sílabo.

· Satorra, A., Ventura, E. (2006), Small-area estimation at idescat: current and related research, IDESCAT.

· Sharma, S. (1996). *Applied Multivariate Techniques*, Willey.

· Spiegelhalter, D. (2006). "Some DIC slides", http://www.mrc-bsu.cam.ac.uk/bugs/WinBUGS/DIC-slides.pdf, acedido em 10 de Outubro de 2010.

· Snijders, T., Bosker, R. (1999), *An Introduction to Basic and Advanced Multilevel Modeling*, Sage Publications.

Anexo I

Constituição do questionário

No próximo domingo vão realizar-se eleições legislativas. Diga qual das seguintes frases se adapta melhor à sua intenção de ir votar?

Gostava agora que me dissesse se considera que o CDS/PP muito, pouco ou nada competente para governar Portugal nos próximos quatro anos?

Qual dos seguintes partidos está em melhores condições para tratar dos problemas do emprego em Portugal?

Em relação ao PS diga se o considera muito, pouco ou nada competente para governar Portugal nos próximos quatro anos?

Qual dos seguintes partidos está em melhores condições para tratar dos problemas da saúde em Portugal?

Em relação ao PSD diga se o considera muito, pouco ou nada competente para governar Portugal nos próximos quatro anos?

Com qual dos seguintes líderes partidários simpatiza mais?

Em relação ao CDU diga se o considera muito, pouco ou nada competente para governar Portugal nos próximos quatro anos?

Em relação ao BE diga se o considera muito, pouco ou nada competente para governar Portugal nos próximos quatro anos?

A probabilidade de ir votar no próximo domingo é?

Em qual dos seguintes partidos pensa votar no próximo domingo?

Antes de terminar, gostaria que me dissesse em qual dos seguintes partidos votou nas eleições legislativas de 2002?

Sexo

Idade

Distrito

Exerce alguma actividade profissional?

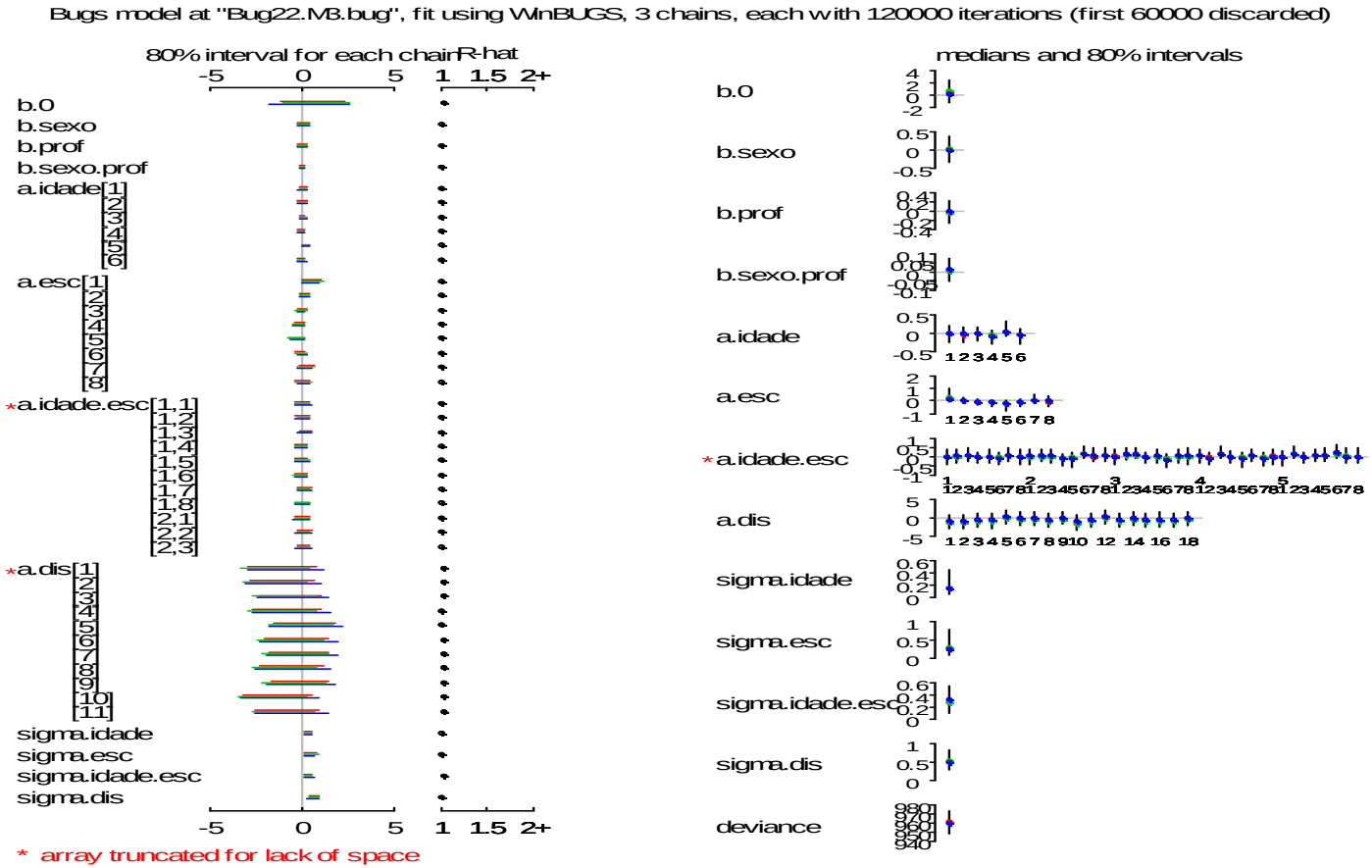
Em qual das seguintes situações se enquadra?

Profissão

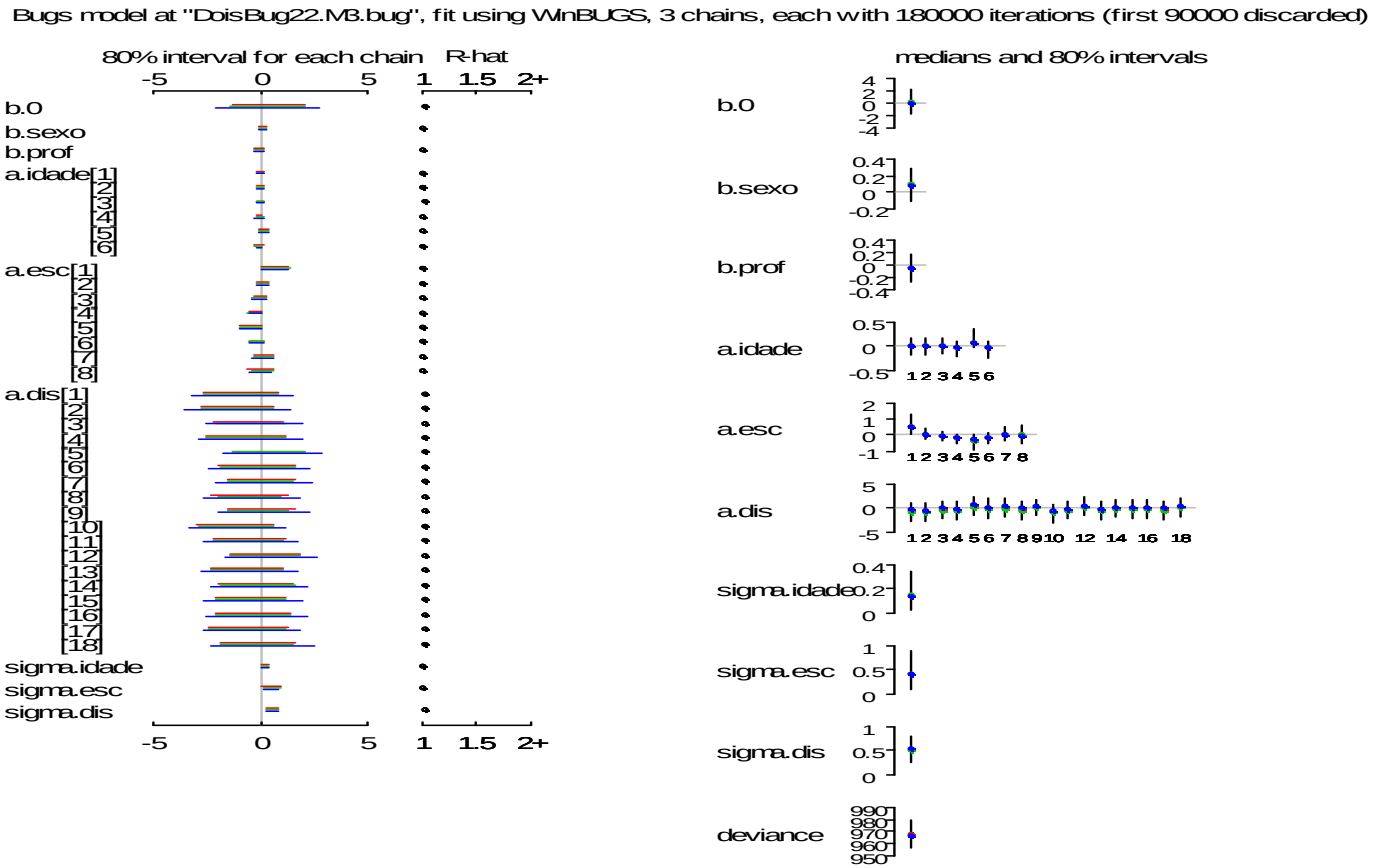
Formação Académica

Anexo II

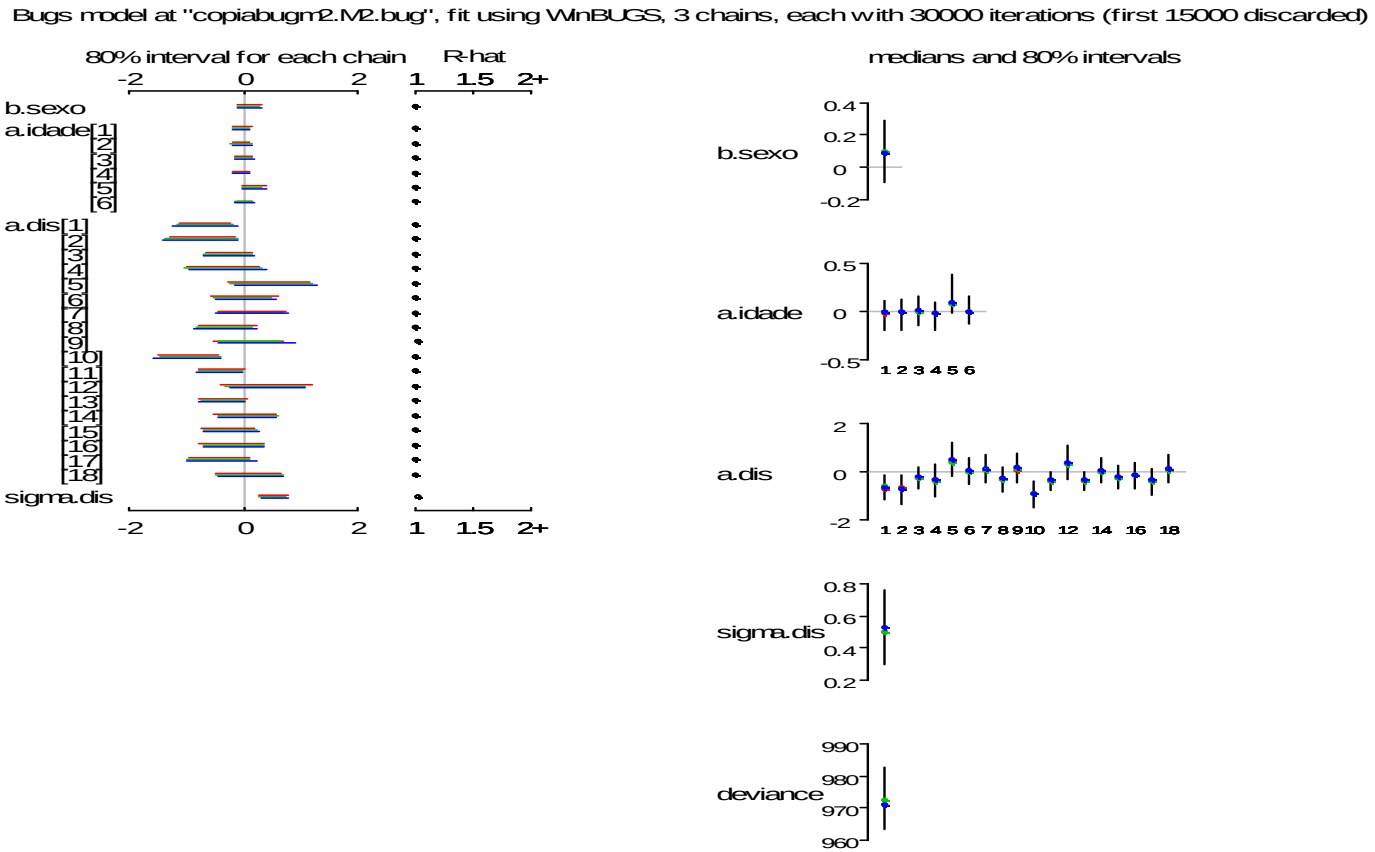
Intervalos para os parâmetros dos modelos, obtidos através do WinBUGS.



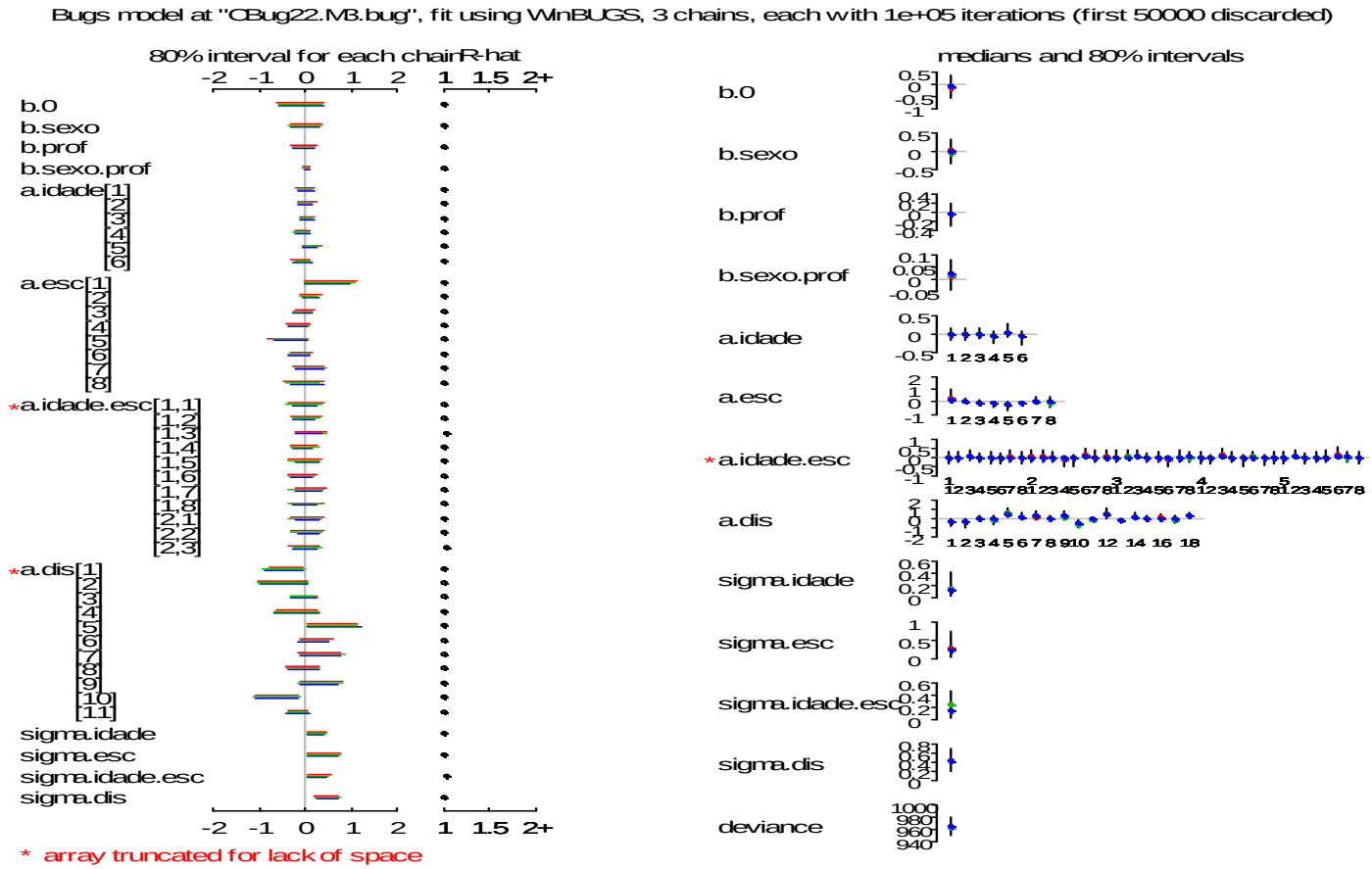
Estimativas parâmetros Modelo 1 obtidas com base nos resultados do WinBUGS.



Estimativas parâmetros Modelo 2 obtidas com base nos resultados do WinBUGS.



Estimativas parâmetros Modelo 3 obtidas com base nos resultados do WinBUGS.



Estimativas parâmetros Modelo 4 obtidas com base nos resultados do WinBUGS.

Anexo III

Código utilizado para a definição dos modelos.

Código R Regressão Multinível

```
dados<-read.csv("base mrp ok.csv", header=T, sep=";")
ine <-read.csv("Ine.csv", header=T, sep=";")
#dadoselei<-read.csv("dadoslei.csv", header=T, sep=";")
dadoselei<-read.csv("dadoslei2.csv", header=T, sep=";")
attach (dados)

set.seed(23)
#seed.save = .Random.seed
#.Random.seed = seed.save

reselei<-dadoselei$X2005
#auxelei<-dadoselei$X2002
auxelei<-dadoselei$media

n.dis<-max(dis)
n.voto<-max(voto)
n.idade<-max(idade)
n.sexo<-max(sexo)
n.esc<-max(esc)
n.prof<-max(prof)
n<-length(voto)

library(arm)

_____
idade.esc <- n.esc*(idade-1) + esc
v.prev <- auxelei[dis]

M1 <- lmer (voto ~ sexo + prof + sexo:prof + v.prev + (1 | idade) + (1 | esc) + (1 |
idade.esc)
+ (1 | dis), family=binomial(link="logit"))
display (M1)
coef(M1)
M1
ranef(M1)

data1 <- list ("voto", "sexo", "idade", "esc", "prof", "dis", "v.prev")
inits1 <- function () {list(set.seed(23),
  b.0=rnorm(1), b.sexo=rnorm(1), b.prof=rnorm(1), b.sexo.prof=rnorm(1),
  a.idade=rnorm(n.idade), a.esc=rnorm(n.esc),
```

```
  a.idade.esc=array (rnorm(n.idade*n.esc), c(n.idade,n.esc)),
  a.dis=rnorm(n.dis),
  sigma.idade=runif(1), sigma.esc=runif(1), sigma.idade.esc=runif(1),
  sigma.dis=runif(1))
}
params1 <- c ("b.0", "b.sexo", "b.prof", "b.sexo.prof",
   "a.idade", "a.esc", "a.idade.esc", "a.dis",
   "sigma.idade", "sigma.esc", "sigma.idade.esc", "sigma.dis")

M1.bugs <- bugs (data1, inits1, params1, "Bug22.M3.bug", n.chains=3, n.iter=120000,
bugs.directory="c:/Programas/WinBUGS14/")

attach.bugs (M1.bugs)
linpred1 <- rep (NA, n)
for (i in 1:n){
  linpred1[i] <- mean (b.0 + b.sexo*sexo[i] + b.prof*prof[i] +
b.sexo.prof*sexo[i]*prof[i]+
a.idade[,idade[i]] + a.esc[,esc[i]] + a.idade.esc[,idade[i],esc[i]])
}



L1 <- nrow (ine)
y.pred1 <- array (NA, c(n.sims, L1))
for (l in 1:L1){
  y.pred1[,l] <- invlogit(b.0 + b.sexo*ine$sexo[l] + b.prof*ine$prof[l]+
b.sexo.prof*ine$sexo[l]*ine$prof[l] + a.idade[,ine$idade[l]] + a.esc[,ine$esc[l]] +
a.idade.esc[,ine$idade[l],ine$esc[l]] + a.dis[,ine$dis[l]])
}

y.pred.dis1 <- array (NA, c(n.sims, n.dis))
for (s in 1:n.sims){
  for (j in 1:n.dis){
    ok <- ine$dis==j
    y.pred.dis1[s,j] <- sum(ine$n[ok]*y.pred1[s,ok])/sum(ine$n[ok])
  }
}

dis.pred1 <- array (NA, c(n.dis,3))
for (j in 1:n.dis){
  dis.pred1[j,] <- quantile (y.pred.dis1[,j], c(.25,.5,.75))
}

dis.pred.med1 <- array (NA, c(n.dis,2))
for (j in 1:n.dis){
  dis.pred.med1[j,1] <- mean(y.pred.dis1[,j])
}

dis.pred.med1 <- array (NA, n.dis)
```

```
for (j in 1:n.dis){
  dis.pred.med1[j] <- mean(y.pred.dis1[,j])
}

mean(abs(dis.pred.med1-reselei))

x<-c(0,0.5,1)

plot(dis.pred.med1,reselei, xlim=c(0.2,0.8), ylim=c(0.2,0.8), xlab="Previsões",
ylab="Valores Reais")
lines(x,y=x, type="l")


sum(dis.pred.med1*dadoselei$N)/sum(dadoselei$N)
detach.bugs(M1.bugs)
```

```
M2 <- lmer (voto ~ sexo + prof + (1|idade)  + (1|esc)
+ (1 | dis)+ v.prev, family=binomial(link="logit"))
display (M2)
coef(M2)
M2
ranef(M2)

data2 <- list ("voto", "sexo", "idade", "esc", "prof", "dis", "v.prev")
inits2 <- function () {list(set.seed(23),
  b.0=rnorm(1), b.sexo=rnorm(1), b.prof=rnorm(1), a.idade=rnorm(n.idade),
a.esc=rnorm(n.esc),
  a.dis=rnorm(n.dis),
  sigma.idade=runif(1), sigma.esc=runif(1), sigma.dis=runif(1))
}
params2 <- c ("b.0", "b.sexo", "b.prof",
   "a.idade", "a.esc", "a.dis",
   "sigma.idade", "sigma.esc", "sigma.dis")

M2.bugs <- bugs (data2, inits2, params2, "DoisBug22.M3.bug", n.chains=3,
n.iter=180000, bugs.directory="c:/Programas/WinBUGS14/")

attach.bugs (M2.bugs)
linpred2 <- rep (NA, n)
for (i in 1:n){
  linpred2[i] <- mean (b.0 + b.sexo*sexo[i] + b.prof*prof[i] +
a.idade[,idade[i]] + a.esc[,esc[i]])
}

L2 <- nrow (ine)
```

```
y.pred2 <- array (NA, c(n.sims, L2))
for (l in 1:L2){
  y.pred2[,l] <- invlogit(b.0 + b.sexo*ine$sexo[l] + b.prof*ine$prof[l]+
a.idade[,ine$idade[l]] + a.esc[,ine$esc[l]] + a.dis[,ine$dis[l]])
}


y.pred.dis2 <- array (NA, c(n.sims, n.dis))
for (s in 1:n.sims){
  for (j in 1:n.dis){
    ok <- ine$dis==j
    y.pred.dis2[s,j] <- sum(ine$n[ok]*y.pred2[s,ok])/sum(ine$n[ok])
  }
}

dis.pred2 <- array (NA, c(n.dis,3))
for (j in 1:n.dis){
  dis.pred2[j,] <- quantile (y.pred.dis2[,j], c(.25,.5,.75))
}

dis.pred.med2 <- array (NA, c(n.dis,2))
for (j in 1:n.dis){
  dis.pred.med2[j,1] <- mean(y.pred.dis2[,j])
}


dis.pred.med2 <- array (NA, n.dis)
for (j in 1:n.dis){
  dis.pred.med2[j] <- mean(y.pred.dis2[,j])
}

mean(abs(dis.pred.med2-reselei))

x<-c(0,0.5,1)
plot(dis.pred.med2,reselei, xlim=c(0.2,0.8), ylim=c(0.2,0.8), xlab="Previsões",
ylab="Valores Reais")
lines(x,y=x, type="l")


sum(dis.pred.med2*dadoselei$N)/sum(dadoselei$N)
___

M3<-lmer(voto ~ sexo+(1|idade)+(1|dis)+v.prev, family=binomial(link="logit"))
display(M3)
coef(M3)
M3
ranef(M3)
```

```
data3 <- list ("voto", "sexo", "idade", "dis", "v.prev")
inits3 <- function () {list(set.seed(23),
  b.sexo=rnorm(1), a.idade=rnorm(n.idade),
  a.dis=rnorm(n.dis),
  sigma.dis=runif(1))
}
params3 <- c ("b.sexo","a.idade","a.dis","sigma.dis")

M3.bugs <- bugs (data3, inits3, params3, "copiabugm2.M2.bug", n.chains=3,
n.iter=30000, bugs.directory="c:/Programas/WinBUGS14/")

attach.bugs (M3.bugs)
linpred3 <- rep (NA, n)
for (i in 1:n){
  linpred3[i] <- mean (a.dis[i] + b.sexo*sexo[i] + a.idade[i])
}

L3 <- nrow (ine)
y.pred3 <- array (NA, c(n.sims, L3))
for (l in 1:L3){
  y.pred3[,l] <- invlogit(a.dis[,ine$dis[l]] + b.sexo*ine$sexo[l] +
    a.idade[,ine$idade[l]])
}

y.pred.dis3 <- array (NA, c(n.sims, n.dis))
for (s in 1:n.sims){
  for (j in 1:n.dis){
    ok <- ine$dis==j
    y.pred.dis3[s,j] <- sum(ine$n[ok]*y.pred3[s,ok])/sum(ine$n[ok])
  }
}

dis.pred3 <- array (NA, c(n.dis,3))
for (j in 1:n.dis){
  dis.pred3[j,] <- quantile (y.pred.dis3[,j], c(.25,.5,.75))
}


dis.pred.med3 <- array (NA, c(n.dis,2))
for (j in 1:n.dis){
  dis.pred.med3[j,1] <- mean(y.pred.dis3[,j])
}

dis.pred.med3 <- array (NA, n.dis)
for (j in 1:n.dis){
  dis.pred.med3[j] <- mean(y.pred.dis3[,j])
}
```

```
x<-c(0,0.5,1)
mean(abs(dis.pred.med3-reselei))
plot(dis.pred.med3,reselei, xlim=c(0.2,0.8), ylim=c(0.2,0.8), main="Modelo 4",
xlab="Previsões", ylab="Valores Reais")
lines(x,y=x, type="l")

sum(dis.pred.med3*dadoselei$N)/sum(dadoselei$N)


___
data4 <- list ("voto", "sexo", "idade", "esc", "prof", "dis", "v.prev")
inits4 <- function () {list(set.seed(23),
  b.0=rnorm(1), b.sexo=rnorm(1), b.prof=rnorm(1), b.sexo.prof=rnorm(1),
  a.idade.raw=rnorm(n.idade), a.esc.raw=rnorm(n.esc),
  a.idade.esc.raw=array (rnorm(n.idade*n.esc), c(n.idade,n.esc)),
  a.dis.raw=rnorm(n.dis),
  sigma.idade.raw=runif(1), sigma.esc.raw=runif(1), sigma.idade.esc.raw=runif(1),
  sigma.dis.raw=runif(1),
  xi.idade=runif(1), xi.esc=runif(1), xi.idade.esc=runif(1), xi.dis=runif(1))
}
params4 <- c ("b.0", "b.sexo", "b.prof", "b.sexo.prof",
   "a.idade", "a.esc", "a.idade.esc", "a.dis",
   "sigma.idade", "sigma.esc", "sigma.idade.esc", "sigma.dis")

M4.bugs <- bugs (data4, inits4, params4, "CBug22.M3.bug", n.chains=3,
n.iter=100000, bugs.directory="c:/Programas/WinBUGS14/")


attach.bugs (M4.bugs)
linpred4 <- rep (NA, n)
for (i in 1:n){
  linpred4[i] <- mean (b.0 + b.sexo*sexo[i] + b.prof*prof[i] +
b.sexo.prof*sexo[i]*prof[i]+
a.idade[,idade[i]] + a.esc[,esc[i]] + a.idade.esc[,idade[i],esc[i]])
}


L4 <- nrow (ine)
y.pred4 <- array (NA, c(n.sims, L4))
for (l in 1:L4){
  y.pred4[,l] <- invlogit(b.0 + b.sexo*ine$sexo[l] + b.prof*ine$prof[l]+
b.sexo.prof*ine$sexo[l]*ine$prof[l] + a.idade[,ine$idade[l]] + a.esc[,ine$esc[l]] +
a.idade.esc[,ine$idade[l],ine$esc[l]] + a.dis[,ine$dis[l]])
}

y.pred.dis4 <- array (NA, c(n.sims, n.dis))
for (s in 1:n.sims){
  for (j in 1:n.dis){
```

```
    ok <- ine$dis==j
    y.pred.dis4[s,j] <- sum(ine$n[ok]*y.pred4[s,ok])/sum(ine$n[ok])
  }
}

dis.pred4 <- array (NA, c(n.dis,3))
for (j in 1:n.dis){
  dis.pred4[j,] <- quantile (y.pred.dis4[,j], c(.25,.5,.75))
}

dis.pred.med4 <- array (NA, c(n.dis,2))
for (j in 1:n.dis){
  dis.pred.med4[j,1] <- mean(y.pred.dis4[,j])
}

dis.pred.med4 <- array (NA, n.dis)
for (j in 1:n.dis){
  dis.pred.med4[j] <- mean(y.pred.dis4[,j])
}

mean(abs(dis.pred.med4-reselei))

x<-c(0,0.5,1)

plot(dis.pred.med4,reselei, xlim=c(0.2,0.8), ylim=c(0.2,0.8), xlab="Previsões",
ylab="Valores Reais")
lines(x,y=x, type="l")


sum(dis.pred.med4*dadoselei$N)/sum(dadoselei$N)
—

#Gráfico
par (mfrow=c(2,2))
plot(dis.pred.med1,reselei, xlim=c(0.2,0.8), ylim=c(0.2,0.8), main="Modelo 1",
xlab="Previsões", ylab="Valores Reais")
lines(x,y=x, type="l")

plot(dis.pred.med2,reselei, xlim=c(0.2,0.8), ylim=c(0.2,0.8), main="Modelo 2",
xlab="Previsões", ylab="Valores Reais")
lines(x,y=x, type="l")

plot(dis.pred.med3,reselei, xlim=c(0.2,0.8), ylim=c(0.2,0.8), main="Modelo 3",
xlab="Previsões", ylab="Valores Reais")
lines(x,y=x, type="l")

plot(dis.pred.med4,reselei, xlim=c(0.2,0.8), ylim=c(0.2,0.8), main="Modelo 3",
xlab="Previsões", ylab="Valores Reais")
```

```
lines(x,y=x, type="l")
```

Código WinBUGS Regressão Multinível

B22.M3.bug

```
".RNG.name" <- "base::Super-Duper"
".RNG.seed" <- 12
model {
  for (i in 1:719){
    voto[i] ~ dbin (p.bound[i], 1)
    p.bound[i] <- max(0, min(1, p[i]))
    logit(p[i]) <- Xbeta[i]
    Xbeta[i] <- b.0 + b.sexo*sexo[i] + b.prof*prof[i] +
      b.sexo.prof*sexo[i]*idade[i] +
      a.idade[idade[i]] + a.esc[esc[i]] + a.idade.esc[idade[i],esc[i]] +
      a.dis[dis[i]]
  }
  b.0 ~ dnorm (0, .0001)
  b.sexo ~ dnorm (0, .0001)
  b.prof ~ dnorm (0, .0001)
  b.sexo.prof ~ dnorm (0, .0001)

  for (j in 1:6) {a.idade[j] ~ dnorm(0, tau.idade)}
  for (j in 1:8) {a.esc[j] ~ dnorm(0, tau.esc)}
  for (j in 1:6) {for (k in 1:8){
    a.idade.esc[j,k] ~ dnorm(0, tau.idade.esc)}}
  for (j in 1:18) {
    a.dis[j] ~ dnorm(a.dis.hat[j], tau.dis)
    a.dis.hat[j] <- b.v.prev*v.prev[j]}
  b.v.prev ~ dnorm (0, .0001)

  tau.idade <- pow(sigma.idade, -2)
  tau.esc <- pow(sigma.esc, -2)
  tau.idade.esc <- pow(sigma.idade.esc, -2)
  tau.dis <- pow(sigma.dis, -2)

  sigma.idade ~ dunif (0, 100)
  sigma.esc ~ dunif (0, 100)
  sigma.idade.esc ~ dunif (0, 100)
  sigma.dis ~ dunif (0, 100)
}
```

Doisbug22.M3

```
".RNG.name" <- "base::Super-Duper"
".RNG.seed" <- 12
model {
  for (i in 1:719){
    voto[i] ~ dbin (p.bound[i], 1)
    p.bound[i] <- max(0, min(1, p[i]))
    logit(p[i]) <- Xbeta[i]
    Xbeta[i] <- b.0 + b.sexo*sexo[i] + b.prof*prof[i] +
      a.idade[idade[i]] + a.esc[esc[i]] +
      a.dis[dis[i]]
  }
  b.0 ~ dnorm (0, .0001)
  b.sexo ~ dnorm (0, .0001)
  b.prof ~ dnorm (0, .0001)

  for (j in 1:6) {a.idade[j] ~ dnorm(0, tau.idade)}
  for (j in 1:8) {a.esc[j] ~ dnorm(0, tau.esc)}
  for (j in 1:18) {
    a.dis[j] ~ dnorm(a.dis.hat[j], tau.dis)
    a.dis.hat[j] <- b.v.prev*v.prev[j]}
  b.v.prev ~ dnorm (0, .0001)

  tau.idade <- pow(sigma.idade, -2)
  tau.esc <- pow(sigma.esc, -2)
  tau.dis <- pow(sigma.dis, -2)

  sigma.idade ~ dunif (0, 100)
  sigma.esc ~ dunif (0, 100)
  sigma.dis ~ dunif (0, 100)
}
```

CopiaBugm2.M2

```
".RNG.name" <- "base::Super-Duper"
".RNG.seed" <- 12
model {
  for (i in 1:719){
    voto[i] ~ dbin (p.bound[i], 1)
    p.bound[i] <- max(0, min(1, p[i]))
    logit(p[i]) <- Xbeta[i]
    Xbeta[i] <- b.sexo*sexo[i] + a.idade[idade[i]] + a.dis[dis[i]]}
  b.sexo ~ dnorm (0, .0001)
  for (j in 1:6) {a.idade[j] ~ dnorm(0, tau.idade)}
  for (j in 1:18) {
    a.dis[j] ~ dnorm(a.dis.hat[j], tau.dis)
    a.dis.hat[j] <- b.v.prev*v.prev[j]}
  b.v.prev ~ dnorm (0, .0001)
  tau.idade<-pow(sigma.idade,-2)
  tau.dis <- pow(sigma.dis, -2)
  sigma.dis ~ dunif (0, 100)
  sigma.idade ~ dunif(0,100)
}
```

CBUG22.M3

```
".RNG.name" <- "base::Super-Duper"
".RNG.seed" <- 12
model {
  for (i in 1:719){
    voto[i] ~ dbin (p.bound[i], 1)
    p.bound[i] <- max(0, min(1, p[i]))
    logit(p[i]) <- Xbeta[i]
    Xbeta[i] <- b.0 + b.sexo*sexo[i] + b.prof*prof[i] +
      b.sexo.prof*sexo[i]*idade[i] +
      a.idade[idade[i]] + a.esc[esc[i]] + a.idade.esc[idade[i],esc[i]] +
      a.dis[dis[i]]
  }
  b.0 ~ dnorm (0, .0001)
  b.sexo ~ dnorm (0, .0001)
  b.prof ~ dnorm (0, .0001)
  b.sexo.prof ~ dnorm (0, .0001)

  for (j in 1:6) {
          a.idade[j] <- xi.idade*(a.idade.raw[j]-mean(a.idade.raw[]))
          a.idade.raw[j] ~ dnorm(0,tau.idade.raw)}
  for (j in 1:8) {
          a.esc[j] <- xi.esc*(a.esc.raw[j]-mean(a.esc.raw[]))
          a.esc.raw[j] ~ dnorm(0,tau.esc.raw)}
  for (j in 1:6) {for (k in 1:8){
```

```
  a.idade.esc[j,k] <- xi.idade.esc*(a.idade.esc.raw[j,k] - mean(a.idade.esc.raw[,]))
  a.idade.esc.raw[j,k] ~ dnorm(0,tau.idade.esc.raw)}}
 for (j in 1:18) {
  a.dis[j] <- xi.dis*(a.dis.raw[j] - mean(a.dis.raw[]))
  a.dis.raw[j] ~ dnorm(a.dis.raw.hat[j],tau.dis.raw)
  a.dis.raw.hat[j] <- b.v.prev.raw*v.prev[j]}

 b.v.prev.raw ~ dnorm(0,.0001)

 tau.idade.raw <- pow(sigma.idade.raw, -2)
 tau.esc.raw <- pow(sigma.esc.raw, -2)
 tau.idade.esc.raw <- pow(sigma.idade.esc.raw, -2)
 tau.dis.raw <- pow(sigma.dis.raw, -2)

 sigma.idade.raw ~ dunif (0, 100)
 sigma.esc.raw ~ dunif (0, 100)
 sigma.idade.esc.raw ~ dunif (0, 100)
 sigma.dis.raw ~ dunif (0, 100)


 xi.idade ~ dunif (0, 100)
 xi.esc ~ dunif (0, 100)
 xi.idade.esc ~ dunif (0, 100)
 xi.dis ~ dunif (0, 100)

 sigma.idade <- xi.idade*sigma.idade.raw
 sigma.esc <- xi.esc*sigma.esc.raw
 sigma.idade.esc <- xi.idade.esc*sigma.idade.esc.raw
 sigma.dis <- xi.dis*sigma.dis.raw

}
```

Código R EBLUP

```
Packages (survey)
Packages (sae2)

# leitura ficheiros
A<-read.csv("A.csv", sep=";")
B<-read.csv("dadoslei.csv", sep=";", dec=".")


# esta linha muda consoante se trate de informação auxiliar de 1999 ou de 2002
dmm<-cbind(data.frame (REG=1:18,DIREST=B$X2002, DESVAR=sd(A$voto),
REGCOV))

dmmeblup<-EBLUP(DIREST~A, ~DESVAR, data=dmm)
dmmeblup
```

Resultados

Para informação auxiliar de 2002:

```
Call:
EBLUP(formula = DIREST ~ A, varformula = ~DESVAR, data = dmm)

Coefficients:
        [,1]
[1,] 0.3067534
[2,] 0.1398341

Variance of the random effects: -0.4968843

Log likelihood: -2134.822
```

Para informação auxiliar de 1999:

```
Call:
EBLUP(formula = DIREST ~ A, varformula = ~DESVAR, data = dmm)

Coefficients:
        [,1]
[1,] 0.3686845
[2,] 0.1479796

Variance of the random effects: -0.4976865

Log likelihood: -3642.842
```

Outros resultados sobre a variância

```
> dmmeblup[]
$coefficients
         [,1]
[1,] 0.3686845
[2,] 0.1479796

$residuals
 [1] -0.0151638953  0.0669055929  0.0049099424 -0.0310763293  0.0280820937
 [6]  0.0173426680 -0.0109375508  0.0492468005  0.0163624492 -0.0351336310
[11] -0.0070652236  0.0147333699  0.0458303939 -0.0004311477 -0.0022800878
[16] -0.0407742873 -0.0220172655 -0.0785338919

$fitted.values
 [1] 0.4168639 0.4003944 0.4381901 0.4278763 0.4889179 0.4540573 0.4673376
 [8] 0.4344532 0.4673376 0.4028336 0.4330652 0.4981666 0.4339696 0.4554311
[15] 0.4387801 0.4426743 0.4296173 0.4605339

$randeff
 [1]   6.3123831 -27.8512695  -2.0438968  12.9363658 -11.6899339  -7.2193564
 [7]   4.5530525 -20.5003177  -6.8113137  14.6253277   2.9410912  -6.1331652
[13] -19.0781457   0.1794770   0.9491485  16.9734041   9.1652845  32.6918647

$varcoeff
      (Intercept)           A
[1,]  0.0006438184 -0.001135726
[2,] -0.0011357260  0.002233941

$varsigma2u
[1] 1.588196e-07

$sigma2u
[1] -0.4976865

$g1
 [1] -207.6732 -207.6732 -207.6732 -207.6732 -207.6732 -207.6732 -207.6732
 [8] -207.6732 -207.6732 -207.6732 -207.6732 -207.6732 -207.6732 -207.6732
[15] -207.6732 -207.6732 -207.6732 -207.6732

$g2
 [1] 24.56504 45.21164 12.14764 16.13541 47.53726 13.39174 21.30882 13.15592
 [9] 21.30882 41.54586 13.65678 63.84279 13.32266 13.92020 12.03378 11.59253
[17] 15.19718 16.47006
```

```
$g3
 [1] 23.13023 23.13023 23.13023 23.13023 23.13023 23.13023 23.13023 23.13023
 [9] 23.13023 23.13023 23.13023 23.13023 23.13023 23.13023 23.13023 23.13023
[17] 23.13023 23.13023

$g1biascorrd
 [1] 1.272709 1.272709 1.272709 1.272709 1.272709 1.272709 1.272709 1.272709
 [9] 1.272709 1.272709 1.272709 1.272709 1.272709 1.272709 1.272709 1.272709
[17] 1.272709 1.272709

$mse
 [1] -138.12040 -117.47379 -150.53780 -146.55003 -115.14818 -149.29369
 [7] -141.37662 -149.52951 -141.37662 -121.13958 -149.02866  -98.84265
[13] -149.36278 -148.76524 -150.65165 -151.09290 -147.48826 -146.21537

$eblup
 [1]   6.7292470 -27.4508751  -1.6057068  13.3642421 -11.2010160  -6.7652991
 [7]   5.0203900 -20.0658645  -6.3439762  15.0281614   3.3741564  -5.6349986
[13] -18.6441761   0.6349081   1.3879286  17.4160784   9.5949018  33.1523985

$method
[1] "ML"

$coef
               [,1]
(Intercept) 0.3686845
A           0.1479796

$rank
[1] 2

$call
EBLUP(formula = DIREST ~ A, varformula = ~DESVAR, data = dmm)

$Z
Matrix of dimension 18x18 with (row-wise) nonzero elements:

 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
Class 'spam'

$desvar
         [,1]
 [1,] 0.498882
 [2,] 0.498882
 [3,] 0.498882
 [4,] 0.498882
 [5,] 0.498882
 [6,] 0.498882
```

[7,] 0.498882
[8,] 0.498882
[9,] 0.498882
[10,] 0.498882
[11,] 0.498882
[12,] 0.498882
[13,] 0.498882
[14,] 0.498882
[15,] 0.498882
[16,] 0.498882
[17,] 0.498882
[18,] 0.498882